Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Biomedica

A.a. 2020/2021

Sessione di Laurea dicembre 2021

Confronto tra ML e DL per la classificazione di alert in radioterapia prodotti dal software PerFraction.

Relatore:

Prof. Filippo Molinari

Correlatore:

Alparone Alessandro

Candidato:

Manuela Contu

Sommario

Introduzione

La radioterapia, utilizzata in più della metà dei pazienti oncologici, persegue l'obiettivo di erogare una dose di radiazioni ionizzanti più alta possibile ai tessuti tumorali causando il minor danno possibile alle cellule e ai tessuti sani.

Se da un lato le tecniche di irradiazione di radioterapia hanno compiuto enormi passi avanti, rendendo possibile l'erogazione di alte dosi terapeutiche, dall'altro lato si è crescentemente forzati a prestare attenzione all'estrema precisione e accuratezza che questi metodi richiedono. Per ottenere processi terapeutici robusti è necessario analizzare incertezze ed errori con l'obiettivo di minimizzarli. Il termine errore in questo caso descrive la deviazione tra trattamento pianificato e trattamento effettivamente erogato.

Alcune cause di incertezza che possono apparire durante il processo di radioterapia sono legate a: diagnosi errata, scarsa qualità dell'imaging in pretrattamento, delineazione del target che varia in base all'interpretazione degli esami o delle linee guida seguite, problemi di erogazione della dose per singole frazioni, per più frazioni o per cause della configurazione del macchinario, correlazioni con il paziente: riduzione o aumento del peso corporeo, riduzione della massa tumorale, movimento naturale degli organi interni, importanti variazioni volumetriche dovute ad un'incorretta preparazione, allineamento del paziente con il fascio laser o utilizzo di sistemi di immobilizzazione errati.

Sofisticati sistemi per la dosimetria in vivo consentono di misurare la dose effettivamente erogata al paziente. Software come PerFraction poi confrontano questa dose con quella pianificata attraverso metriche specifiche come analisi gamma, istogrammi dose-volume o calcoli puntuali di differenza di dose. Se il software rileva in queste metriche valori al di fuori di specifiche soglie di accettabilità impostate dagli operatori del reparto di radioterapia di ogni centro, restituisce un alert, ovvero comunica al team di radioterapia che la dose effettivamente erogata al paziente in quella seduta si è discostata dalla distribuzione di dose definita in fase di pianificazione.

Il software PerFraction è in grado di notare una variazione tra la dose di riferimento e quella misurata, ma non è in grado di dire se questo scostamento sia dovuto ad un errore di erogazione da parte della macchina, un errore dovuto a uno scorretto setup del paziente sul lettino di trattamento o ancora un errore dovuto a variazioni anatomiche del paziente.

L'obiettivo dello studio è stato quello di elaborare un confronto di performance tra algoritmi di ML, un SVM e un Multilayer Perceptron, allenati su dati quantitativi di dose e algoritmi di DL, applicando la

tecnica di transfer learning ad una GoogleNet e ad una ResNet, allenati su immagini RGB formate dall'unione di immagini CT e immagini CBCT. I dati alla base dello studio sono stati concessi dal reparto di radioterapia dell'Istituto IRCC di Candiolo in collaborazione con l'Azienda Tecnologie Avanzate S.R.L.

Nello specifico, trattandosi di un progetto molto complesso, si è partiti dall'analisi di un solo distretto anatomico, il distretto della prostata, presentato brevemente al Capitolo 1. Successivamente nel Capitolo 2 sono spiegati i principi fisici alla base della radioterapia, i dispositivi necessari all'erogazione della stessa ma anche all'imaging e al monitoraggio, le diverse tecniche di erogazione più utilizzate e infine gli step seguiti in un percorso radioterapico dal momento della diagnosi fino al completamento del trattamento. Nel Capitolo 3 è spiegata l'utilità di sistemi QA (quality assurance) per il paziente, le caratteristiche del software PerFraction, i dispositivi per la dosimetria in vivo e le metriche per il controllo giornaliero di dose. Nel Capitolo 4 sono elencati materiali e metodi selezionati per lo svolgimento dello studio e nel Capitolo 5 sono riportati i risultati ottenuti con discussione e commento degli stessi.

1. Tumore della prostata

La prostata, presente solo negli uomini, è una piccola ghiandola situata nella zona pelvica, tra il pene e la vescica, davanti al retto. Ha le dimensioni di una noce, lunga circa 3 cm, larga 4 cm e spessa 2,5 cm. La sua principale funzione è quella di produrre il liquido prostatico che costituisce circa un quarto dello sperma totale [1].

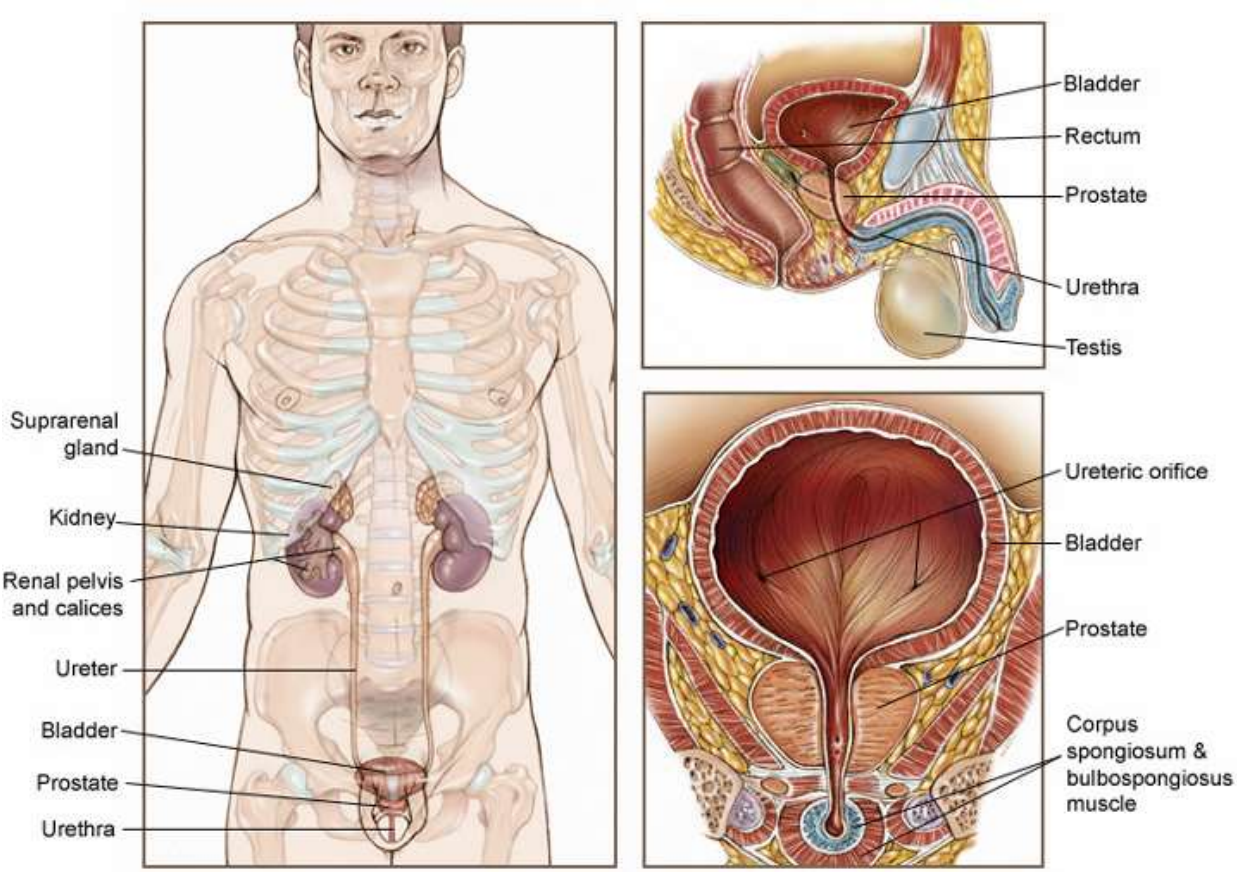

Figura 1 Rappresentazione anatomica dell'apparato genitale maschile, e della ghiandola prostatica. [1]

Questa ghiandola può essere bersaglio di diverse patologie, tra cui neoplasie, per le quali l'età risulta essere il più importante fattore di rischio. Il cancro alla prostata, infatti, è raro sotto i 40 anni e sempre più frequente con l'aumentare dell'età, circa il 60% dei casi sono diagnosticati in uomini over 65, l'età media è di 66 anni. La percentuale di sopravvivenza ai 5 anni (ovvero la percentuale di persone sopravvissute per almeno 5 anni dopo la diagnosi di tumore), per il tumore alla prostata è uguale a quella dei 10 anni, pari al 98%. La percentuale di sopravvivenza ai 5 anni, per pazienti con cancro locale o regionale (la diagnosi avviene quando il tumore è ancora localizzato nella sola prostata o al più in strutture molto vicine), è del 100%. Questa percentuale invece si abbassa drasticamente al 30% per uomini in cui la diagnosi avviene quando il tumore si è già diffuso in altre parti del corpo [1].

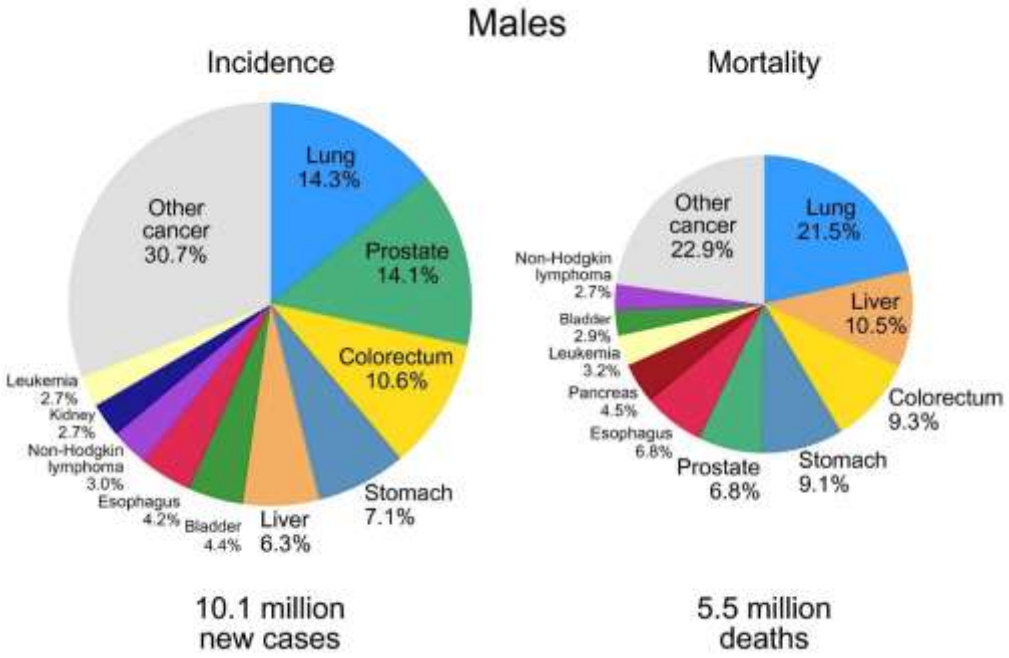

Figura 2 Distribuzione di incidenza e morti per i 10 tumori più comuni nel 2020 per gli uomini. Le aree del diagramma a torta riflettono la proporzione del numero totale di casi o morti. I tumori della pelle sono inclusi nella sezione 'other cancers' [2].

Incidenza e mortalità mondiale per tumori maschili nel 2020 sono mostrati in Figura 2. Il tumore alla prostata, da solo rappresenta il 14,1% delle diagnosi, risultando come la seconda patologia più frequente. Con il 6,8% su 5.5 milioni di morti, questo tipo di tumore si posiziona quinto nel diagramma delle mortalità [2].

Il carcinoma della prostata tende a crescere lentamente negli anni, per questa ragione è spesso asintomatico nei suoi stadi iniziali. Alcuni sintomi sono fastidi urinando, sangue nelle urine, problemi di erezione, dolore a schiena, anche, costole e altre ossa. Quando più sintomi suggeriscono questo tipo di patologia, diversi test diagnostici possono essere eseguiti:

- Esame del PSA: L'antigene prostatico specifico (PSA) è una proteina prodotta dalle cellule (sia sane che tumorali) della ghiandola prostatica. Il PSA si trova principalmente nello sperma ma piccole quantità possono anche essere presenti nel sangue. Quando i livelli di PSA salgono, sale di pari passo la possibilità di avere un tumore alla prostata, anche se non esiste una soglia specifica che aiuti la diagnosi [3].
- Biopsia prostatica: la biopsia prostatica è necessaria quando i risultati di altri test diagnostici (PSA, esame digito rettale (DRE) o altri) suggeriscono la presenza di un tumore. La biopsia è una procedura in cui un campione della prostata è rimosso per permettere l'analisi al microscopio. Eseguita da un urologo, la core needle biopsy è il metodo principale per la diagnosi di questo tipo di carcinoma [3].

Di seguito si elencano altri test che possono essere eseguiti in contemporanea o in aggiunta ai precedenti per migliorare l'accuratezza della diagnosi o della stadiazione del tumore in una fase successiva.

- Ecografia prostatica trans rettale (TRUS): consiste in una piccola sonda inserita per un breve tratto nel retto permettendo di ottenere immagini anatomiche della ghiandola prostatica. Viene eseguita a seguito di un PSA elevato o un DRE anomalo o per visualizzare la prostata durante una biopsia, mentre ha una bassa affidabilità diagnostica nell'identificare le neoplasie [3].
- MRI: come nel caso precedente l'imaging a risonanza magnetica (MRI) non ha valenza diagnostica ma può essere utilizzato durante la biopsia per indirizzare correttamente l'ago verso la prostata oppure per determinare lo stadio tumorale o ancora per verificare se il tumore si è diffuso al di fuori della prostata, nelle vesciche seminali o in altre strutture vicine.
- Tomografia computerizzata (CT): la tomografia computerizzata utilizza raggi X per generare immagini dettagliate del corpo. Un'immagine CT può mostrare se il tumore si è espanso al di fuori della prostata [3].
- Tomografia ad emissione di positroni (PET): anche questo test può essere utile per tracciare l'espansione del tumore. Mentre altri test come raggi X, CT o MRI rivelano cambiamenti anatomici, la PET è utilizzata per mostrare cambiamenti chimici e fisiologici. Prima dell'esecuzione della PET vera e propria una piccola quantità di sostanza contenente uno zucchero attaccato ad un isotopo radioattivo è iniettato nel sangue del paziente. Le cellule tumorali assorbono lo zucchero con l'isotopo che emette positroni (radiazioni a bassa energia). I positroni reagiscono con gli elettroni delle cellule tumorali producendo raggi gamma, poi rilevati dal macchinario PET che traduce questa informazione in immagine. Questo test può anche essere eseguito in contemporanea a un MRI (PET-MRI) o a una CT (PET-CT) [3].

Una volta eseguita la diagnosi e accertata la presenza del tumore si passa alla fase di terapia. Gli approcci più comuni sono la chemioterapia (CT), la radioterapia (RT), la chirurgia e l'immunoterapia. Questi metodi possono essere più o meno efficaci se usati singolarmente o in combinazione [4].

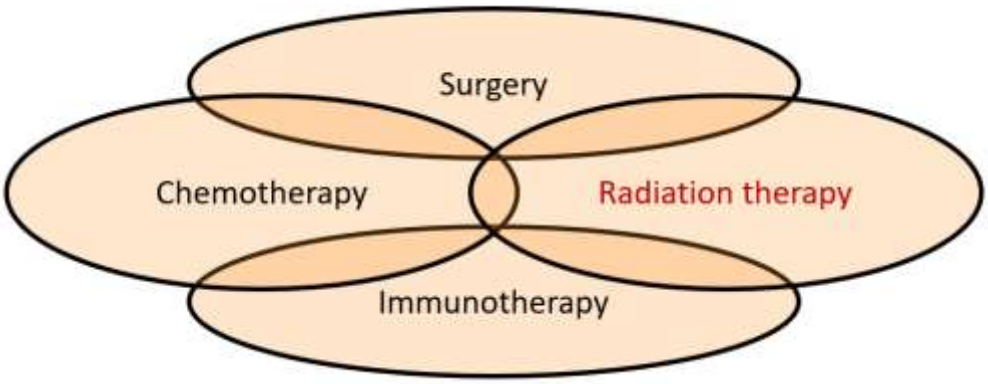

Figura 3 Principali metodiche di cura tumorale, per monoterapia o approccio multimodale [4].

La radioterapia, in base allo stadio tumorale e ad altri fattori, può essere utilizzata:

- Come primo trattamento per un tumore circoscritto nella ghiandola prostatica e ad uno stadio iniziale.
- Come primo trattamento ma in combinazione con una terapia ormonale per tumori già espansi al di fuori della prostata e in tessuti circostanti.
- Come secondo trattamento se, in seguito alla chirurgia, il tumore non è stato rimosso completamente o se nonostante la totale rimozione si ripresenta.
- Come ultimo trattamento, per un tumore in stadio avanzato, per cercare di tenere la massa tumorale sotto controllo, per prevenire o alleviare i sintomi [4].

2. Radioterapia

La radioterapia è quella branca specialistica della medicina che si avvale dell'uso di radiazioni ionizzanti nel trattamento delle malattie neoplastiche. Si può partire proprio da questa definizione elementare per capire nel dettaglio la fisica delle radiazioni, le applicazioni e i vantaggi. Ma prima di inoltrarci nel campo della radioterapia può essere utile soffermarci su due termini specifici presenti nella definizione iniziale: malattie neoplastiche. Una neoplasia non è altro che un processo patologico che causa una proliferazione incontrollata delle cellule di un tessuto. Queste cellule proliferanti possono rendersi più o meno indipendenti dai processi di differenziazione e accrescimento dei tessuti, portando alla crescita di masse abnormi di tessuto distruggendo le normali strutture anatomiche provocando spesso la perdita progressiva della funzione del tessuto o organo coinvolto. Il tessuto neoplastico generalmente si differenzia dal tessuto originale e questa caratteristica viene utilizzata a nostro vantaggio durante l'erogazione della radioterapia. Le radiazioni ionizzanti, dirette contro la massa tumorale, e nello specifico verso il suo DNA, generano radicali liberi che danneggiano irreparabilmente il patrimonio genetico cellulare, danneggiando quindi la struttura del tessuto bersaglio; inoltre, le cellule tumorali sono scarsamente capaci di riparare i propri danni e quindi una volta colpite dalle radiazioni vanno incontro a morte cellulare o apoptosi [5][6].

La radioterapia, utilizzata in più della metà dei pazienti oncologici, persegue l'obiettivo di erogare una dose di radiazioni più alta possibile ai tessuti tumorali causando il minor danno possibile alle cellule e ai tessuti sani. L'intenzione, quindi, non è solo di distruggere le cellule tumorali e prolungare la vita del paziente ma anche di fare in modo che la vita stessa sia di alta qualità [7] [8].

2.1 Fisica delle radiazioni

Fin dalla nascita della radiologia, la fisica delle radiazioni ne è stata una parte integrante. È sull'interazione fisica tra radiazioni, sia elettromagnetiche che corpuscolari, e atomi che costituiscono la materia che si radica l'essenza della radiologia diagnostica e terapeutica. La radiazione può essere vista come un mezzo per trasferire energia da una sorgente a un oggetto distante.

La produzione di raggi X di interesse clinico avviene quando elettroni con energia cinetica compresa tra 10keV e 50meV sono decelerati da specifici target metallici. Quando questo avviene la gran parte dell'energia cinetica degli elettroni viene dispersa in calore nel target, mentre una piccola parte di energia viene emessa sotto forma di fotoni a raggi X comunemente divisibili in due gruppi: i raggi X caratteristici e i raggi X bremsstrahlung. I primi sono generati dall'interazione di coulomb tra gli elettroni incidenti gli elettroni orbitali del materiale target. I secondi sono il risultato di interazioni Coulombiane tra gli elettroni incidenti e il nucleo del materiale target.

I raggi X possono essere usati sia in radiologia oncologica o radioterapia per la cura tumorale, sia in radiologia diagnostica per la diagnosi di patologie. I tubi a raggi X producono raggi X superficiali (con energia cinetica degli elettroni tra 10keV e 100keV) e raggi X orthovoltage (con energia cinetica degli elettroni tra 100keV e 500keV) mentre i LINAC (acceleratori lineari) producono i raggi X megavoltage (con energia cinetica degli elettroni superiore a 1MeV). Di seguito verrà elencata nel dettaglio tutta l'attrezzatura utilizzata in ambito radioterapico [9].

2.2 Macchinari per la radioterapia a fasci esterni

La terapia e fasci esterni (EBT dall'inglese external beam therapy) è una tecnica di radioterapia oncologica che si contraddistingue da radioterapia interna e brachiterapia proprio per la localizzazione della fonte radioattiva, esterna al corpo umano. L'EBT permette di erogare al paziente un fascio di dose generalmente generato da un acceleratore lineare. Solitamente, prima dell'erogazione della terapia, il paziente viene sottoposto a uno scan TC che permette l'individuazione di localizzazione e forma della massa tumorale.

Di seguito verranno elencati i macchinari solitamente utilizzati durante il processo di erogazione di radioterapia a fasci esterni, con particolare focus sui dispositivi presenti nell'Istituto IRCC di Candiolo, con il quale si è collaborato per l'acquisizione dei dati necessari allo svolgimento dello studio.

2.2.1 Acceleratore lineare

Figure 4 LINAC- acceleratore lineare per l'erogazione di radioterapia [4].

I LINAC medici, ovvero acceleratori lineari, prendono il loro nome dalla caratteristica principale che li contraddistingue: accelerano gli elettroni fino al raggiungimento di energie cinetiche che variano da 4MeV a 25MeV usando campi a microonde non conservativi nel range di frequenza da 10^3 MHz a 10^4 MHz.

La Figura 5 riporta uno schema della struttura del LINAC, ci sono importanti differenze tra macchine di aziende diverse ma le parti fondamentali per la generazione del fascio di elettroni caratterizzanti questo macchinario sono:

- Sistema di iniezione (*electron gun*)
- Generatore di potenza RF (*RF power generator*)
- Guida d'onda (*accelerating waveguide*)
- Sistema di trasporto del fascio (*electron beam transport*)
- Collimatore e sistema di monitoraggio [9] [10].

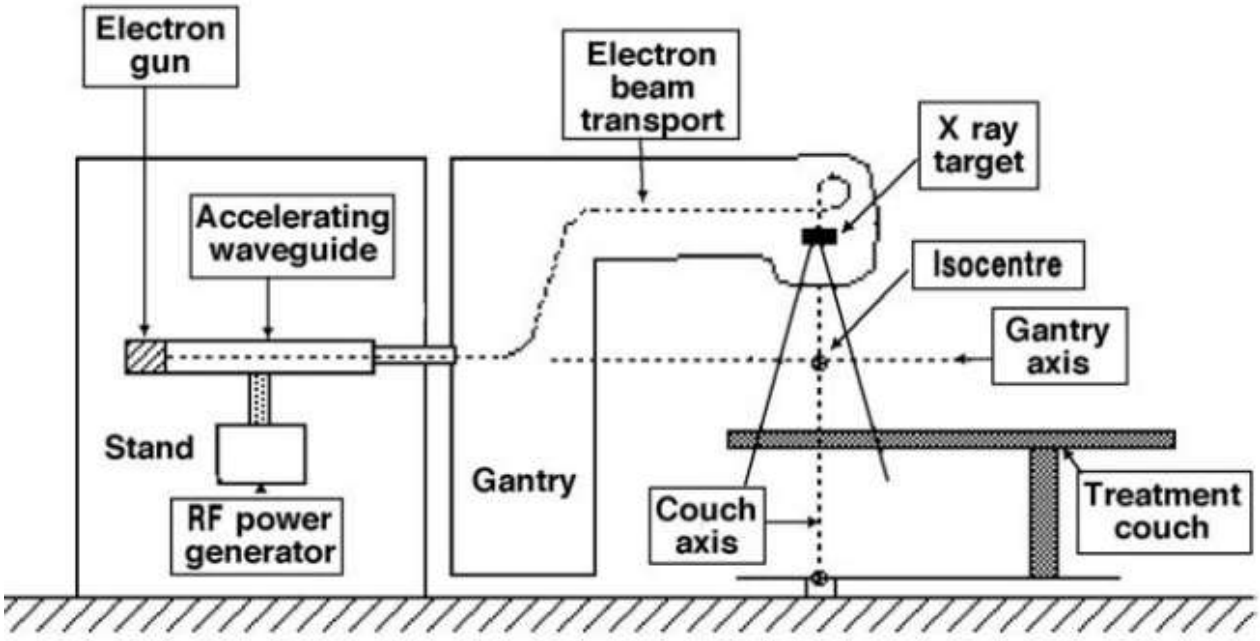

Figura 5 Rappresentazione schematica di un LINAC isocentrico al medico [9].

Gli elettroni sono prodotti dal sistema di iniezione, *electron gun,* un semplice acceleratore elettrostatico composto da un catodo (filamento riscaldato) e un anodo forato messo a terra; la differenza di potenziale che si crea tra i due accelera gli elettroni prodotti verso la guida d'onda. A questo punto, il generatore di potenza ha il compito di produrre campi magnetici a radiofrequenza che accelerino gli elettroni fino al raggiungimento dell'energia cinetica desiderata. Gli elettroni, trasportati attraverso un condotto (*electron beam transport*) finiscono per collidere contro un target metallico producendo quindi raggi X. Un collimatore multilamellare (MLC) a questo punto ha il compito di

adattare il fascio a raggi X ad alta energia, in uscita dalla macchina, alla geometria del tumore e direzionarlo. Il paziente soggetto alla terapia è posizionato sul lettino mobile sotto il gantry, il quale può essere fatto ruotare attorno al paziente per irradiare da diverse angolazioni la massa tumorale [9] [10].

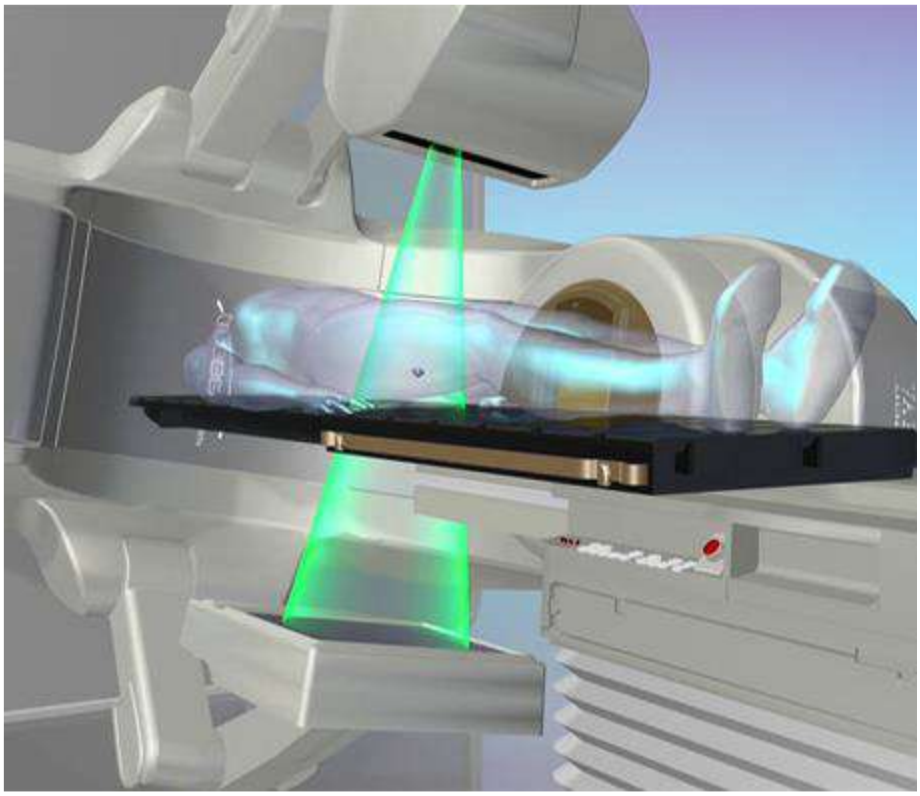

Figura 6 CBCT, integrato nella struttura del LINAC, usato per l'acquisizione di immagini necessarie al setup del paziente prima dell'erogazione della terapia [4].

L'implementazione della radioterapia guidata dalle immagini (IGRT) è stata facilitata dall'avvento di moderne modalità di imaging volumetrico, spesso integrate nel LINAC stesso. In figura 6 si vede un sistema di imaging fornito di funzione CBCT incorporato nella struttura di sostegno del gantry [11].

2.2.2 Cone-Beam Computed Tomography (CBCT)

La CBCT sviluppata negli anni '90 è entrata in uso nell'ambito radioterapico negli anni 2000. Questo sistema di imaging utilizza un fascio di radiazioni in MVoltage erogato dal LINAC (come in Figura 6), o un fascio in kVoltage erogato usando un tubo a raggi X addizionale, montato sul LINAC. La CBCT è un vero e proprio strumento di IGRT per la verifica del posizionamento del paziente in sede di terapia [11].

La CBCT consente al tecnico radiologo di correggere cambiamenti nella posizione del volume target prima del trattamento, ma anche di monitorare la preparazione del paziente, variazioni nell'anatomia del paziente o nella morfologia del tumore [12]. I benefici introdotti dalla CBCT hanno portato la

radioterapia guidata da immagini ad essere una tecnica di routine utilizzata su scala globale. Sono diverse le aziende che hanno integrato la CBCT all'unità di erogazione della radioterapia: Varian (utilizzata all'Istituto di Candiolo con cui si è collaborato per lo svolgimento dello studio. Il True Beam LINAC rappresentato in figura 5 è di proprietà Varian), Elekta, Siemens e Vero.

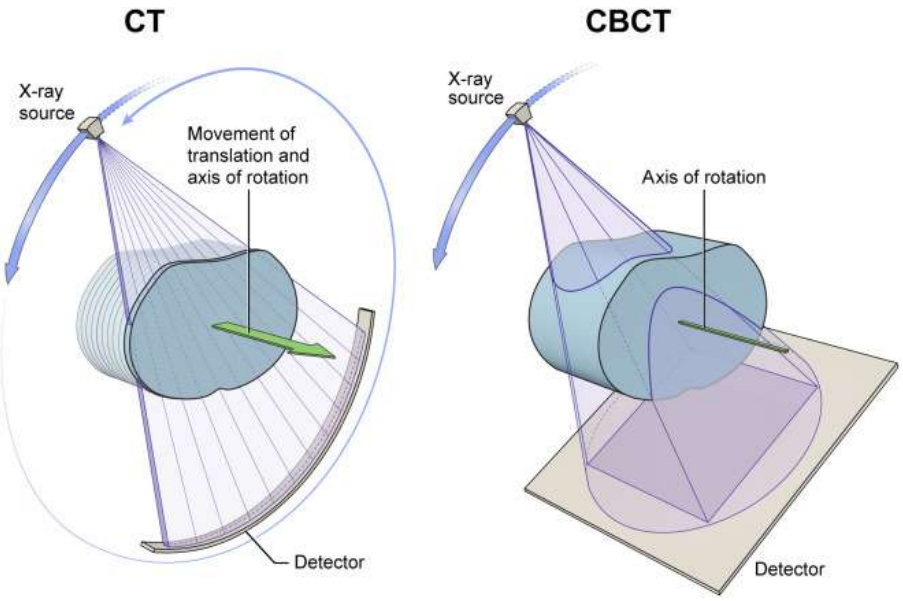

Figura 7 Confronto CT (a sinistra) e CBCT (a destra). La CBCT ha una sola sorgente di raggi X che eroga un ampio fascio conico durante una singola rotazione. La CT è composta da diversi fasci a ventaglio che irradiano una piccola area, il che richiede rotazioni multiple della sorgente durante il movimento del paziente [13].

Nella tecnica di imaging CBCT si utilizza un fascio a raggi X di forma piramidale o conica che compie un'unica rotazione, contemporaneamente al rilevatore, di 360° attorno al paziente acquisendo una serie di immagini bidimensionali utilizzate in un secondo momento per ricostruire il volume di interesse con un algoritmo di retroproiezione filtrata. Per contro la CT è caratterizzata da uno stretto fascio a raggi X, a forma di ventaglio (fan beam) e da un gruppo di rilevatori. Il paziente in questo caso, non rimane fermo come accade per la CBCT, ma deve essere mosso continuamente mentre il gantry gli ruota attorno.

Il principale vantaggio nell'utilizzo della CBCT si trova nel ridotto tempo di acquisizione che si traduce anche in una diminuzione di esposizione del paziente alle radiazioni.

2.2.3 Electronic Portal Imaging Device (EPID)

L'EPID è un sistema di rilevazione di radiazione posizionato sotto il gantry del sistema di erogazione della terapia per catturare le radiazioni in uscita dal paziente. L'EPID è anche utilizzato per studi sulla garanzia di qualità della componente di imaging, senza il paziente sul lettino di trattamento, e da

sistemi come SunCheck (presentato al Capitolo 3) come un sistema di verifica di dose prima del trattamento o in vivo.

All'Istituto IRCC di Candiolo l'EPID aSi1000 è integrato nel LINAC TrueBeam di Varian. Questo rilevatore è montato sul supporto robotico chiamato braccio ad E (per la sua forma), perpendicolare al gantry e può essere regolato a una distanza che varia da 95 a 180 cm dalla sorgente di radiazione. L'area di imaging dell'EPID è di 1024*767 pixel con una risoluzione di 0.39mm e una capacità di ottenere immagini a 14bit in 30fps(frames/secondo). Le modalità di acquisizione di immagini con questo dispositivo sono: prima, durante o dopo il trattamento. Oltre ad essere usato per monitorare le frazioni di trattamento può anche essere usato per scopi dosimetrici, come sarà spiegato successivamente nel Capitolo 3.

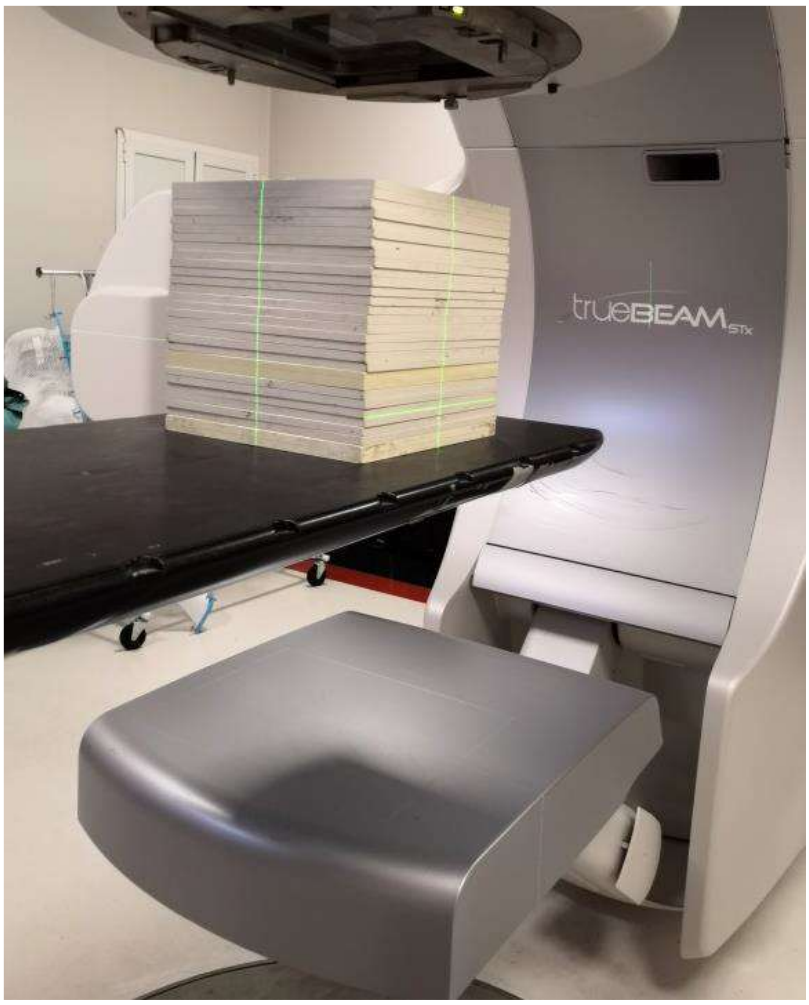

Figura 8 Sistema EPID integrato con il LINAC True Beam [4].

2.3 Tecniche di erogazione della radiazione

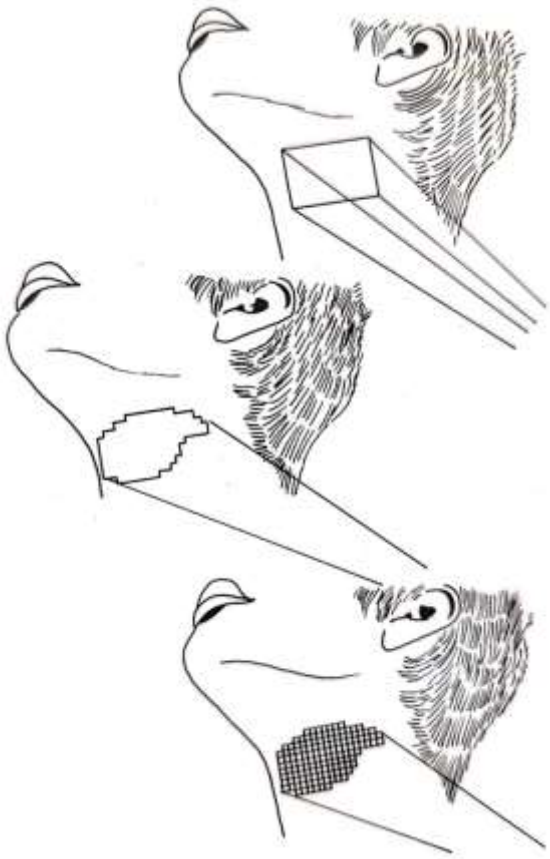

Figura 9 Radioterapia convenzionale (in alto), radioterapia conformazionale (CFRT) senza modulazione di intensità (in mezzo), CFRT con modulazione di intensità (IMRT)(in basso) [14].

In Figura 9 sono mostrate le differenze chiave tra radioterapia convenzionale, radioterapia conformazionale (CFRT) senza modulazione di intensità e con modulazione di intensità. Durante il primo secolo di esistenza della radioterapia è stato possibile erogare solo campi di forma rettangolare (radioterapia convenzionale). Con l'arrivo del collimatore multilamellare (MLC) si sono potuti elaborare campi con forme geometriche più strategiche (CFRT). La continua evoluzione dell'apparecchiatura per radioterapia ha permesso di migliorare la conformazione della distribuzione di dose attorno al volume target con una conseguente riduzione dell'irradiazione dei tessuti sani. L'IMRT è la forma più avanzata di CFRT in quanto permette non solo di modellare geometricamente il fascio ma anche di variare pixel per pixel l'intensità all'interno del campo di dose. Questa caratteristica risulta particolarmente vantaggiosa quando il volume target presenta una superficie concava o è giustapposto ad organi a rischio. Ad esempio, in Figura 9 si mostra il distretto testa-collo dove il tumore può essere adiacente a colonna vertebrale, occhi, nervo ottico e ghiandole parotidee [14].

2.3.1 Radioterapia conformazionale con modulazione di intensità (IMRT)

La radioterapia con modulazione di intensità rappresenta una delle maggiori innovazioni tecniche nella radioterapia moderna. L'IMRT è un trattamento conformazionale tridimensionale che utilizza schemi con fasci ad intensità non uniforme con un sistema di ottimizzazione computerizzato per ottenere una distribuzione di dose più meticolosa. Grazie alla possibilità di gestire le intensità dei singoli raggi del fascio, l'IMRT consente un ottimo controllo della distribuzione di dose che in combinazione con tecniche immagini-guidate per delineare con precisione i volumi target ed erogare il piano di cura, può migliorare il controllo tumorale riducendo tossicità ai tessuti sani. Anche forti cadute di dose sul confine tra target e organi a rischio possono essere possibili grazie alla modellazione di distribuzioni di dose complesse. Questa tecnologia ha permesso di ridurre notevolmente il volume di strutture critiche irradiate da alte dosi portando a notevoli miglioramenti nei risultati [15].

2.3.2 Volumetric Modulated Arc Therapy (VMAT)

La VMAT è una delle tecniche più moderne nell'ambito del trattamento dei tumori e, grazie al notevole miglioramento di efficienza di erogazione rispetto al campo statico nell'IMRT, ha attirato molto l'attenzione nell'ambito medico. È in grado di erogare la dose dinamicamente mentre il gantry è in movimento permettendo un grande numero di direzioni dei fasci a differenza dell'IMRT che tipicamente permette meno di dieci angolature per fasci a campo fisso. Le lamelle del MLC cambiano continuamente posizione al variare dell'angolazione del gantry per soddisfare le variazioni di morfologia del target. L'erogazione di un'alta dose in un tempo relativamente corto è quindi possibile grazie a questa tecnica di erogazione dinamica in cui intensità di dose e forma del campo di dose sono modificati in continuo. La tecnica VMAT è applicabile a tutte le neoplasie, ma tenendo in mente le considerazioni appena fatte, risulta evidente come il suo uso sia vantaggioso per quelle patologie in cui è necessario ottimizzare al massimo la precisione di erogazione sul target tumorale per preservare organi a rischio molto vicini. Alcuni casi più comuni sono i tumori testa-collo, tumori toracici (seno e polmoni), tumori pelvici come cancro al retto o alla prostata [16].

2.4 Elementi di radioterapia oncologica

Comprensione di tipologia, estensione della malattia da curare e considerazioni sugli effetti del trattamento su tessuti e organi sani sono necessari per un attento utilizzo delle radiazioni ionizzanti. Il lavoro d'equipe è un requisito fondamentale per il trattamento di pazienti oncologici. Il team di cura, in un trattamento di radioterapia, consiste in oncologo radioterapista, dosimetrista, tecnico di radioterapia e fisico medico [17].

Dopo che la diagnosi di tumore è stata confermata dalla biopsia l'oncologo di radiazione valuta il caso clinico del paziente. Definito poi lo stadio tumorale, se si è deciso di optare per un percorso di

radioterapia, si prosegue con la pianificazione del trattamento, il calcolo di dose e infine l'erogazione della terapia. Non meno importanti delle fasi di pianificazione sono poi i controlli periodici del paziente durante le sedute e al termine del percorso radioterapico [17].

2.4.1 Stadiazione

La comprensione della storia naturale di ogni malattia è alla base del trattamento per la cura. La stadiazione è un modo per descrivere in maniera schematica quanto è grande un tumore e quanto si è esteso rispetto alla sede originale di sviluppo.

La classificazione del tumore avviene in seguito alla valutazione del tumore primario, dei linfonodi circostanti e di eventuali metastasi. Nel momento della diagnosi il tumore può essere in uno stadio precoce, intermedio o avanzato [17].

2.4.2 Decisione terapeutica

Dopo la fase di stadiazione si prosegue con la scelta del percorso clinico da intraprendere. Si parla di trattamento curativo se ci sono possibilità di guarigione del paziente oncologico. Invece, nel caso non ci siano speranze di eradicare totalmente il tumore, si opta per un trattamento palliativo con l'intento di ridurre la sofferenza e allungare la speranza di vita del paziente. Se in questa fase si decide di intraprendere un percorso di radioterapia seguiranno la pianificazione della terapia e il calcolo della dose da erogare [17].

2.4.3 Pianificazione del trattamento

La pianificazione è un processo molto delicato durante il quale vengono definiti: volume di trattamento, dose totale destinata alla massa tumorale, numero di frazioni, dose per frazione e frequenza delle frazioni.

La definizione del volume tumorale e del volume di trattamento (target volume) sono requisiti essenziali per la pianificazione del trattamento radioterapico. Nei trattamenti curativi il volume tumorale è dato da tutta la massa maligna. Il volume di trattamento viene sempre considerato più ampio del volume tumorale per diverse ragioni: coprire estensioni microscopiche del tumore difficili da individuare, permettere un margine di errore durante la definizione del volume tumorale, e per ragioni più pratiche quali ammettere variazioni dovute a movimenti come la respirazione del paziente durante il trattamento. In questa fase, l'oncologo ha il compito di individuare, oltre alla massa tumorale, anche tutte le strutture (organi e tessuti) nella zona di trattamento, facendo un contouring delle stesse sulla CT di pianificazione. A questo punto il fisico sanitario può ottimizzare il piano di radioterapia tenendo anche in considerazione gli effetti su tessuti e organi che la radiazione incontra

durante il suo percorso verso il tumore. Infatti, esistono organi e tessuti, essenziali per la sopravvivenza funzionale del paziente, che hanno limiti di dose massima accettabile [17].

2.4.4 Trattamento

Dopo l'approvazione del piano da parte dell'oncologo il paziente può iniziare il trattamento, percorso durante il quale sarà assistito dal tecnico di radioterapia. Durante il trattamento è richiesto uno sforzo giornaliero da parte del paziente per minimizzare le variazioni tra la condizione di pianificazione e la condizione di erogazione quotidiana [17].

2.5 Incertezze ed errori in radioterapia

Ogni step del trattamento di radioterapia è caratterizzato da un certo livello di incertezza, è impensabile di eseguire la procedura perfettamente. Per determinare le incertezze di dose erogata a tessuti umani, sani o tumorali, è necessario individuare e comprendere le incertezze associate alle diverse fasi del processo radioterapico.

Quantitativamente, l'errore (o deviazione) di una misura o di un calcolo è definito come la differenza tra il suo valore e il valore atteso, ovvero un valore considerato come riferimento e ottenuto tramite altri metodi. La dispersione dei valori ottenuti per una specifica misura eseguita periodicamente invece identifica il parametro dell'incertezza.

Se da un lato le tecniche di irradiazione di radioterapia hanno compiuto enormi passi avanti, rendendo possibile l'erogazione di alte dosi terapeutiche, dall'altro lato si è crescentemente forzati a prestare attenzione all'estrema precisione e accuratezza che questi metodi richiedono. Una verifica più attenta sul posizionamento di volume target e organi a rischio è richiesta a causa dell'altro livello di dose conformazionale raggiungibile con le tecniche moderne. Ottenere un'ottima corrispondenza tra dose calcolata e dose erogata sul volume target è uno dei principali obiettivi in radioterapia. Per raggiungere un processo robusto bisogna analizzare possibili incertezze ed errori per cercare di minimizzarli. Il termine errore in questo caso è utilizzato per descrivere la deviazione tra trattamento pianificato e trattamento effettivamente erogato. L'analisi delle incertezze in radioterapia ha un ruolo fondamentale per salvaguardare la qualità della vita del paziente. Conseguenza significative sul paziente come paralisi o danni irreversibili possono avvenire in caso di importanti errori durante l'erogazione della dose. Siccome le incertezze non devono mai essere ignorate, diverse strategie possono essere introdotte per monitorare i possibili errori e far sì che se presenti abbiano un impatto minimo sul trattamento radioterapico.

Di seguito si riportano le cause di alcune incertezze che possono apparire durante il processo di radioterapia:

- Diagnosi errata: interpretazione errata degli esami istologici
- Imaging: la scarsa qualità dell'imaging in pretrattamento
- Contouring del volume target: la delineazione del target può essere differente in base all'interpretazione degli esami o delle linee guida seguite.
- Prescrizione della dose: incertezze sui limiti o sulle tolleranze per alcuni organi a rischio
- Erogazione della terapia: problemi di erogazione della dose per singole frazioni, per più frazioni o per cause della configurazione del macchinario.
- Correlazioni con il paziente:
    - Anatomia: riduzione o aumento del peso corporeo, riduzione della massa tumorale, movimento naturale degli organi interni, importanti variazioni volumetriche dovute ad un'incorretta preparazione.
    - Setup: mancato allineamento del paziente con il fascio laser, utilizzo di sistemi di immobilizzazione errati

In radioterapia si è soliti parlare di clinical target volume (CTV) e planning target volume (PTV). Il CTV è il volume reale della massa tumorale da trattare, il PTV invece è un'espansione del CTV, un volume più ampio che garantisce margini di sicurezza. L'introduzione del PTV è un metodo per ridurre l'impatto degli errori di setup, i più comuni in radioterapia e garantire una completa copertura di trattamento del tumore. Altri errori possibili sono quelli legati al movimento degli organi e della massa tumorale dovuti a respirazione, pulsazione delle arterie o peristalsi. Infine, gli errori anatomici, meno frequenti, ma non meno importanti possono portare a una mancata corrispondenza tra la determinazione morfologica del tumore in fase di pianificazione e quella in fase di trattamento. Il risultato terapeutico può essere compromesso da queste variazioni che possono modificare le correlazioni tra diversi tessuti, sia sani che patologici. Questo aspetto ha una particolare importanza quando si utilizza la tecnica VMAT, caratterizzata da gradienti di dose altamente conformi ai margini del PTV, per cui anche variazioni anatomiche o di posizionamento modeste possono rendere il piano di cura non ottimale. In queste condizioni si può verificare un sotto dosaggio del volume tumorale o un sovradosaggio di strutture e organi a rischio. Queste situazioni possono accadere quotidianamente ma specialmente durante le ultime sedute quando i tessuti hanno assorbito un'elevata dose e i fenomeni di tossicità sono più frequenti. Per i motivi sopra elencati risulta necessario mettere in atto dei sistemi di garanzia di qualità specifici per il paziente che monitorino la dose erogata quotidianamente al paziente.

3. PerFraction - Garanzia di qualità paziente specifica

Una formazione adeguata del personale, programmi di garanzia della qualità (QA) appropriati, strumenti e procedure di controllo qualità (QC) sono fondamentali per garantire livelli adeguati di precisione e accuratezza in radioterapia. Una buona procedura per la riduzione del rischio clinico deve includere [18]:

- Procedure che riguardino tutti i processi chiave nell'organizzazione
- Monitoraggio dei processi
- Procedure di conservazione della documentazione
- Registrazione di incidenti avvenuti o sfiorati, e conseguenti azioni correttive messe in atto
- Revisioni regolari dei processi e del sistema di qualità stesso

Solitamente un programma QA completo è composto da un programma QA macchina, che verifichi e dimostri che le caratteristiche del macchinario non deviino significativamente da valori di riferimento acquisiti al momento del collaudo, e un programma QA paziente-specifico. L'obiettivo principale di un QA paziente è quello di garantire che l'impatto del trattamento sul paziente, dovuto sia alle performance globali del macchinario che a fattori umani, sia coerente con il piano di cura. Un QA paziente è composto da un QA pretrattamento e dalla dosimetria in vivo. Il primo permette di verificare che il piano di cura erogato sia confrontabile con quello pianificato. La dosimetria in vivo invece serve a verificare che tutto il processo di delivery del trattamento sia riproducibile e confrontabile con la fase di pianificazione (incluso setup del paziente, preparazione al trattamento, erogazione macchina, ecc).

Di seguito si riporta un esempio concreto di software, PerFraction (PF) di SunNuclear, utilizzato in ambito ospedaliero per soddisfare i QA paziente: PF è la piattaforma alla base dello sviluppo di questa tesi.

SunNuclear è stata fondata nel 1984 come una società fornitrice di servizi di manutenzione e calibrazione, per poi crescere nel tempo fino a diventare un leader rispettato nella progettazione e produzione di sistemi di qualità per misurazioni di radiazioni. SunCheck di SunNuclear è una piattaforma software destinata a raccogliere, individuare, confrontare, calcolare, analizzare, mostrare e conservare dati di dosimetria e QA in radioterapia. Più precisamente, SunCheck è un'applicazione web, basata su un server, con una piattaforma che integra QA paziente, QA macchina e workflow di gestione dati, accessibile da qualsiasi PC collegato in rete.

Patient QA	PlanCHECK
	DoseCHECK
	PerFRACTION
Machine QA	SNC Machine
	SNC Routine

Tabella 1 Struttura della piattaforma SunCheck composta dai due moduli Patient QA e Machine QA. Nella colonna di destra i sotto moduli.

PerFraction è il sottomodulo del QA paziente che si occupa dell'analisi della dosimetria in vivo. PF è progettata per permettere l'individuazione di errori che possono accadere durante l'erogazione di un trattamento di radioterapia; è uno strumento che permette di effettuare QA per ogni frazione di un piano di radioterapia, inclusa la così detta frazione 0, ovvero il pretrattamento. Questo software utilizza informazioni rilevate dal fascio di dose in uscita dal paziente per confrontare le caratteristiche di trattamento tra diverse frazioni rispetto a quelle attese dal piano di cura, fornendo una verifica consistente e giornaliera sull'erogazione del piano. Questo confronto permette di individuare gli errori che possono accadere per colpa dell'anatomia o del setup del paziente, o per ragioni legate al sistema di erogazione come il MLC, l'angolo di rotazione del gantry, la dose in uscita o all'apertura del collimatore. PF automatizza tutti i requisiti del QA paziente, dalla verifica secondaria del piano di trattamento al QA pretrattamento e al monitoraggio in vivo, usando immagini EPID (Electronic Portal Imaging Device) e/o Log-files (file generati dall'acceleratore e contenenti informazioni riguardanti parametri geometrici e dosimetrici di ogni piano di trattamento erogato dallo stesso).

In questo modo, attraverso una sola applicazione web tutti i dati e i risultati sono conservati nel Database SunCheck. PF è principalmente diviso in tre parti, corrispondenti alle 3 fasi critiche di un piano di trattamento:

- Verifica secondaria della dose – DoseCheck: effettua calcoli secondari 3D della dose per le tecniche conformazionali (3DCFRT) e volumetriche (IMRT, VMAT), per macchine differenti (LINAC, Tomoterapia, Brachiterapia).
- QA PerFraction per il pretrattamento – frazione 0: QA durante la fase di pre-trattamento usando l'EPID e/o i dati ottenuti dai Log File (che consentono un analisi 3D), o l'EPID da solo per un'analisi 2D indipendente.
- QA PerFraction in vivo – frazione N: permette il monitoraggio giornaliero della dose ricevuta dal paziente durante l'erogazione della terapia usando l'EPID e/o i dati ottenuti dai Log File (che consentono un'analisi 3D), o l'EPID da solo per un'analisi 2D indipendente. Analizza la

dose durante il trattamento per individuare eventuali possibili errori associati alla macchina o al paziente.

Siccome il focus della tesi è quello di andare a sviluppare un classificatore che automatizzi l'interpretazione degli errori individuati da PerFraction durante l'erogazione della radioterapia si analizza solo l'ultima delle tre fasi sopra descritte. Iniziamo facendo chiarezza sulle differenze tra analisi 2D e 3D di cui si è appena parlato.

3.1 Confronto generale tra analisi 2D e 3D

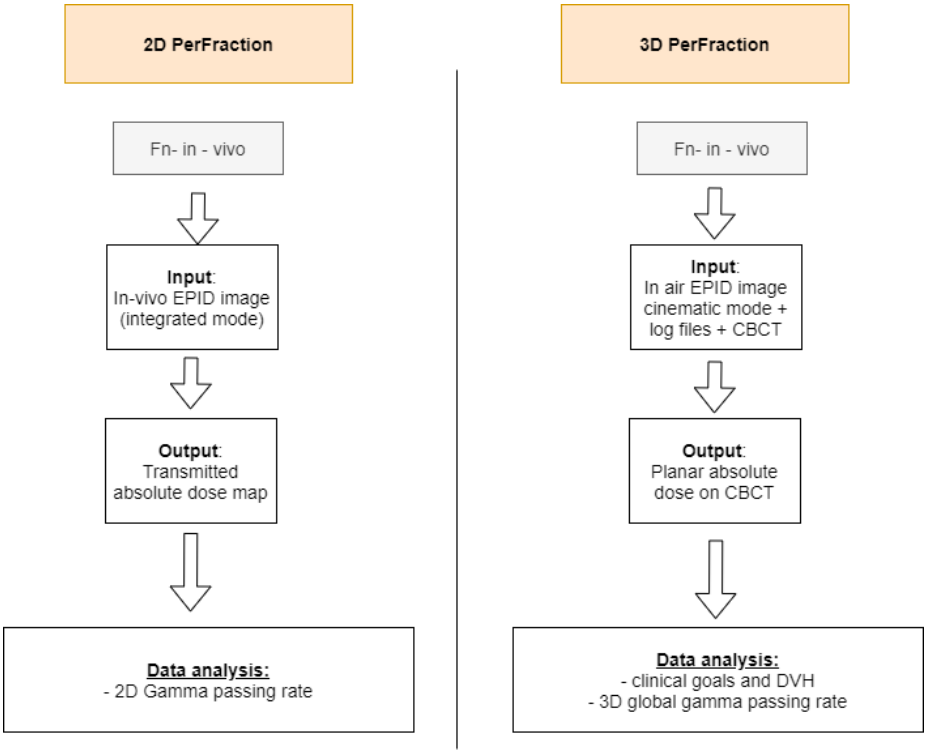

Figura 10 Confronto schematico tra le modalità 2D (sinistra) e 3D (destra) di PerFraction. Si possono osservare i dati necessari in input, quelli restituiti come output e le analisi eseguite sui dati [4]

La Figura 10 rappresenta schematicamente le differenze chiave tra analisi 2D e 3D. L'analisi 2D converte l'immagine EPID in una mappa assoluta di dose, poi la confronta con la mappa di dose di transito attesa ottenuta dal piano di radioterapia (piano RT). Il confronto dei risultati è effettuato con una metrica gamma (2D) per ogni fascio di erogazione.

L'analisi 3D esegue una ricostruzione di dose basata sulle informazioni estratte dal collimatore multilamellare (MLC), sulle posizioni del collimatore ottenute analizzando le immagini EPID e sulle informazioni ricavate dai log files macchina (come angolazione del gantry e unità monitor (MU)). Un algoritmo di convolution/superposition è utilizzato per ricalcolare, a partire dalla fluenza ricostruita, con queste informazioni la distribuzione di dose sulla CBCT per il monitoring in vivo giornaliero.

3.2 Dosimetria in vivo (IVD)

In primis definiamo il significato di dosimetria in vivo. 'In vivo' deriva dal latino e significa 'nel vivente', ovvero caratterizza studi e prove effettuate su un organismo vivente. In opposizione alle misurazioni di dose 'ex vivo' o 'in vitro' (fatte prima o dopo il trattamento usando un fantoccio al posto del paziente), la IVD è la misurazione della dose, ricevuta dal paziente durante il trattamento, attraverso un dosimetro che può essere posizionato esternamente al paziente. Nella maggior parte dei casi il dosimetro è posizionato sulla pelle o nelle vicinanze del distretto anatomico irradiato e la sua risposta viene poi correlata con la dose all'interno del paziente usando relazioni specifiche tra la dose misurata e informazioni anatomiche del paziente. Molte ricerche nel tempo hanno mostrato come programmi di calcolo della dose indipendenti o misurazioni pretrattamento non sono stati efficaci per l'identificazione di un gran numero di errori accaduti durante il trattamento, come lo sarebbero state le misurazioni di dose in vivo. La dosimetria in vivo, in EBRT, ha quindi principalmente l'obiettivo di individuare errori significativi e di verificare la corrispondenza tra dose erogata al paziente giorno per giorno e dose pianificata. Nella sezione successiva andremo ad analizzare le tecnologie attualmente disponibili per gli studi di IVD. Il dosimetro, strumento usato per misurare una dose assorbita di radiazioni ionizzanti, insieme al suo specifico lettore costituisce il sistema di dosimetria. Affinché un dosimetro sia utile deve avere: alta precisione, alta accuratezza, ampio range in dinamica, alta risoluzione spaziale, ridotta dipendenza spaziale, ridotta dipendenza dalla dose. Storicamente, la dosimetria in vivo prevedeva l'utilizzo di rivelatori (diodi o mofset) posizionati sulla cute del paziente. Tale sistema di misura di tipo puntuale risultò essere limitato con l'introduzione delle tecniche volumetriche, poiché la misura di dose in un punto non risultava più rappresentativa di tutto il trattamento.

3.2.1 Analisi dei Log Files

Il log file non è altro che un file elettronico contenente le registrazioni dei parametri dinamici dei macchinari per l'erogazione della radioterapia come: posizione delle lamelle del MLC, posizione del collimatore, posizione del gantry, unità monitor. L'analisi tramite log file usa una combinazione di log files e algoritmi per il calcolo della dose per stimare la dose erogata al paziente durante ogni trattamento. In questo caso non si ha una misura diretta della dose ma un ricalcolo a partire dai dati registrati giornalmente dal LINAC.

3.2.2 Dosimetria con EPID

Tutti i LINACS sono equipaggiati con l'EPID, ideato inizialmente per verificare il posizionamento del paziente, e poi utilizzato anche per misurare la dose erogata giornalmente grazie alle sue caratteristiche positive come: rapida acquisizione di immagini (centinaia di ms), alta risoluzione spaziale (sub-millimetrica), possibilità di ottenere immediatamente il risultato in formato digitale e

larghe aree di rilevazione. La dosimetria tramite EPID, ormai completamente automatizzata, può essere svolta giornalmente su ogni fascio di ogni paziente [19].

Generalmente le procedure di dosimetria con EPID si dividono in misurazioni:

- Pre - trattamento: procedura per il confronto del piano di trattamento con le misurazioni del fascio di radiazione erogato dal LINAC prima del trattamento, quindi in assenza di paziente.
- Durante il trattamento: confronto tra la distribuzione di dose erogata e il piano di trattamento basato su misurazioni effettuate durante lo svolgimento di una seduta di radioterapia, quindi ovviamente in presenza del paziente.

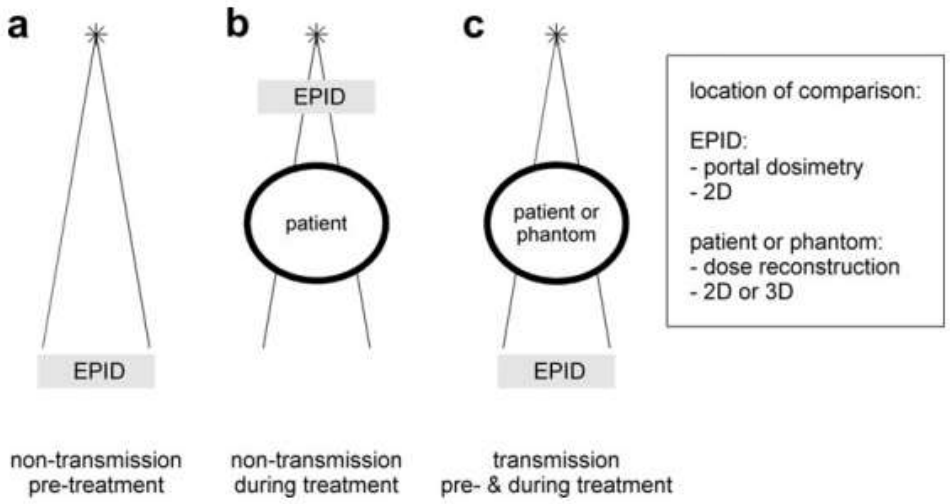

Figura 11 Tre disposizioni per la dosimetria con EPID [20].

In modo generico però, indipendentemente dal tipo di dosimetro scelto, le metodologie per svolgere dosimetria possono essere distinte in relazione alla posizione del dosimetro rispetto al paziente (o fantoccio) [20].

- Dosimetria di non trasmissione: il rilevatore misura la dose erogata al paziente (o fantoccio) o la fluenza dell'energia incidente essendo posizionato tra la sorgente e il paziente stesso.
- Dosimetria di trasmissione: il rilevatore misura la dose erogata al paziente (o fantoccio) o la fluenza dell'energia incidente trasmessa attraverso il paziente (o fantoccio) posizionato in mezzo tra sorgente e dosimetro.

Questo studio in particolare si occuperà di dosimetria in vivo, di tipo trasmissivo, utilizzando l'EPID come rivelatore.

Negli anni sono stati svolti molti studi per ottimizzare il confronto tra immagini EPID di transito con un'immagine di dose di riferimento (al livello dell'EPID), e per ricostruire la distribuzione di dose

all'interno del paziente. Una IVD basata su EPID può avvenire in due modalità: dosimetria 2D planare o dosimetria volumetrica 3D. Nella prima, proiezione per il confronto di immagini di trasmissione 2D, si predice un'immagine di dose dal piano di trattamento e dalle immagini CT (usato per la pianificazione della cura) e la si paragona all'immagine di dose misurata durante la seduta specifica.

Nella seconda, una tecnica di retroproiezione per la ricostruzione da immagini di trasmissione, si ha la combinazione dell'immagine EPID di dose misurata con un algoritmo di retroproiezione per calcolare la distribuzione di dose in ogni voxel della CT o della CBCT giornaliera ricevuto dal paziente durante la frazione di radioterapia.

3.3 Metriche

Il positivo sviluppo di tecnologie di erogazione sofisticate come l'IMRT o la VMAT ha portato con sé, non solo benefici ma anche la necessità di migliorare di pari passo programmi per un'analisi accurata della qualità della dose erogata durante i trattamenti. Infatti, il confronto tra distribuzione di dose pianificata e distribuzione di dose erogata è un requisito fondamentale per il successo di un percorso radioterapico. Al fine di soddisfare queste necessità sono state sviluppati metodi di confronto qualitativi e quantitativi. DI seguito si presentano le metriche più comunemente usate per valutare la dosimetria in vivo. In radioterapia si ricorre spesso all'indice gamma (ϒ) e all'istogramma dose-volume. In aggiunta a queste metriche, per completare l'analisi dosimetrica, solitamente si osserva anche un valore puntuale: la dose all'isocentro.

3.3.1 Gamma index

Questa metrica è una delle più utilizzate per la verifica della qualità del trattamento di radioterapia erogato con tecniche come l'IMRT e la VMAT. Gran parte dei software commerciali di analisi del processo radioterapico implementano l'analisi gamma (ϒ), che combinando differenza di dose e distanza fornisce un mezzo molto valido per il confronto tra dose erogata e calcolata.

3.3.1.1 Definizione di ϒ

L'indice gamma consente di valutare la distribuzione di dose punto per punto attraverso una metrica adimensionale che combina differenze di dose e di distanza. Lo standard a cui confrontare la distribuzione di dose valutata giornalmente, è solitamente la dose di pianificazione, o dose di riferimento che può essere un singolo punto (misurato con una camera di ionizzazione), un profilo lineare (1D), un profilo 2D o 3D. La distribuzione di dose da valutare è solitamente quella predetta dal TPS [21].

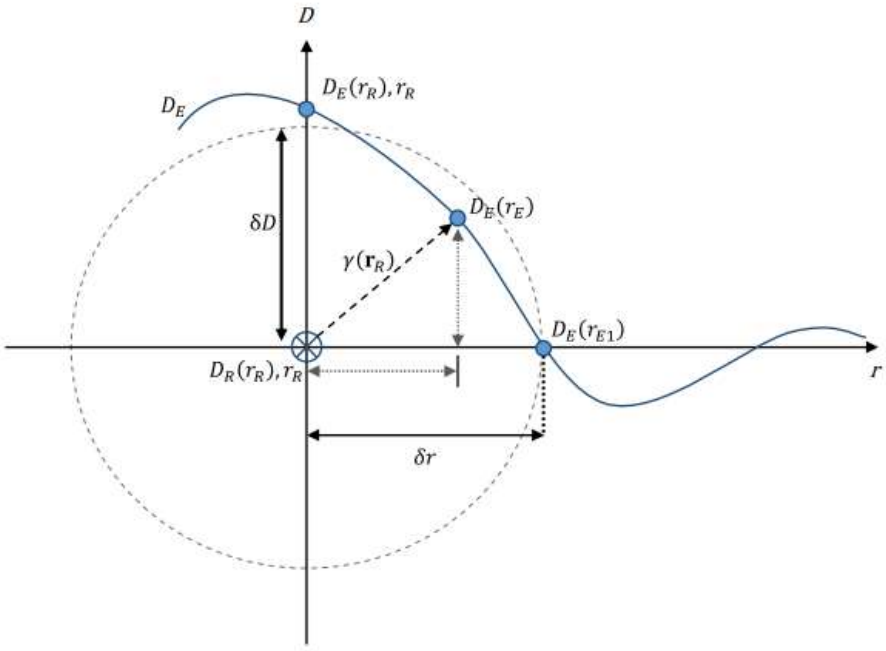

Figura 12 Rappresentazione schematica del metodo gamma 1D. Sull'asse y c'è la dose D, sull'asse x la distanza r. L'origine è il punto di riferimento r_R e la linea blu rappresenta la distribuzione di dose analizzata. I criteri δr e δD formano un'ellisse che racchiude i punti accettabili intorno ad r_R. [21].

Per calcolare l'indice ϒ è necessario calcolare, per ogni punto nella distribuzione di dose:

- La distanza tra il punto di riferimento e quello da valutare $\Delta r(r_R, r_E)$
- La differenza di dose tra il punto di riferimento e quello da valutare $\Delta D(r_R, r_E)$.

Dove r_R è il punto di riferimento e r_E è il punto da valutare. La differenza di dose è calcolata come

$\Delta D(r_R, r_E)=D_E(r_E)- D_R(r_R)$ dove $D_E(r_E)$ è la dose in un punto della distribuzione di dose da valutare e $D_R(r_R)$ un punto della distribuzione di dose di riferimento. Dopodiché per ogni punto della distribuzione giornaliera da valutare si calcola:

$$\Gamma(r_R, r_E)= \sqrt{\frac{\Delta r^2(r_R,r_E)}{\delta r^2} + \frac{\Delta D^2(r_R,r_E)}{\delta D^2}}$$

Dove δr è il criterio per la differenza di distanza e δD è il criterio per la differenza di dose. Infine, la ϒ si ottiene come il minimo valore ottenuto tra tutti i punti considerati.

$$\Upsilon(r_R)= \min \{ \Gamma(r_R, r_E)\} \forall \{r_E\}$$

Come si vede da Fig. 12, i criteri δD e δr formano un elissoide intorno al punto di riferimento. Se un punto da valutare è localizzato all'interno di quest'elissoide allora il punto supera l'analisi gamma in quanto ϒ<1. È standard comune riportare i criteri nel formato $\delta D(\%)/\delta r(\mathrm{mm})$. Il criterio più usato è

3%/3mm ma, come si noterà nel Capitolo 4, per lo sviluppo di questo studio sono stati usati anche altri 3 criteri: 2%/3mm, 5%/3mm, 10%,3mm [21].

3.3.1.2 Calcolo locale e globale di ϒ

Per la metrica ϒ è possibile effettuare sia un calcolo globale sia un calcolo locale. La differenza sta semplicemente nella dose di riferimento considerata. Nella ϒ locale $\Delta D(r_R, r_E)=D_E(r_E)- D_R(r_R)$ mentre nella ϒ globale $\Delta D(r_R, r_E)= \frac{(D_E(r_E)- D_R(r_R))}{D_{norm}}$ in cui D_{norm} è un valore di dose qualsiasi scelto per la normalizzazione (solitamente il valore di D_{max} del piano di cura). Entrambe le tipologie di analisi gamma, locale e globale, hanno vantaggi e svantaggi. Nell'analisi ϒ locale, per ogni punto di dose misurato, la tolleranza di dose è calcolata sulla base della dose locale di quel punto, all'interno dei limiti di distanza. Per questo in zone a bassa dose e in zone con alto gradiente di dose è più probabile che l'intorno locale non soddisfi i criteri gamma, per questo motivo la metrica risulta troppo restrittiva. La scelta sulla tipologia di metodica dipende strettamente dalle necessità specifiche per ogni valutazione. Come si vedrà in seguito, nella parte di sviluppo della tesi, per l'analisi della dosimetria sono state usate sia la gamma locale sia quella globale [21].

3.3.2 Dose Volume Histogram (DVH)

Gli istogrammi dose-volume sono uno strumento molto funzionale per una valutazione quantitativa della distribuzione di dose erogata durante i trattamenti mentre sono poco funzionali per una valutazione spaziale della distribuzione stessa. Molto semplicemente un DVH è una rappresentazione grafica della distribuzione in frequenza dei valori di dose all'interno di un volume definito, che sia il volume target o una struttura anatomica qualsiasi.

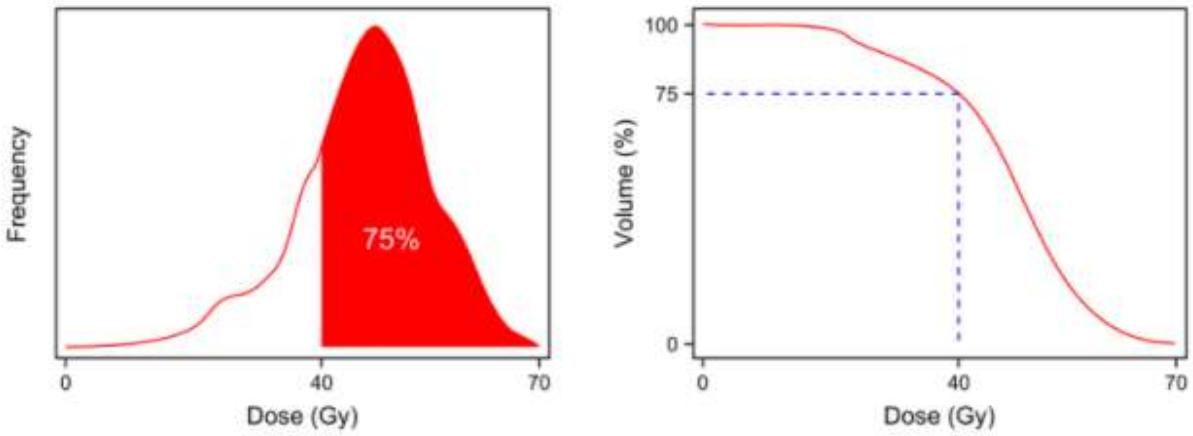

Figura 13 DVH differenziale (a sinistra) e cumulativo (a destra). In questo caso ad esempio si nota come il 75% del volume totale riceva una quantità di dose maggiore o uguale a 40Gy [22].

Esistono due tipologie di DVH: differenziale o cumulativo, entrambe rappresentate in Figura 13. Il primo rappresenta per ogni valore di dose (asse x) la percentuale di volume totale (asse y) che ha ricevuto quella specifica quantità di dose. Il secondo invece rappresenta sull'asse y la percentuale di volume totale che ha ricevuto una quantità di dose uguale o maggiore del corrispondente valore sull'asse delle ascisse.

Per valutare efficacemente la dose effettivamente erogata al paziente in fase di trattamento è possibile individuare, a partire dalla distribuzione di dose volumetrica (voxel per voxel), dei punti di dose del DVH che forniscano informazioni quantitative sulla copertura del volume target. Questo tipo di analisi consente di identificare eventuali sotto dosaggi o sovradosaggi di specifiche strutture. Le statistiche più comunemente utilizzate sono:

- Dose di copertura del volume (D95%= dose minima che copre il 95% del volume totale)
- Dose massima nel volume (D2%= dose erogata solamente al 2% del volume totale)
- Dose media nel volume (D50%= dose media erogata alla metà del volume totale)

Queste tre metriche, applicate al target, saranno utilizzate nello sviluppo pratico della tesi.

3.3.3 Dose puntuale composita

Solitamente questa metrica viene utilizzata per calcolare la differenza puntuale tra il valore di dose all'isocentro dalla distribuzione di dose del TPS e quello della distribuzione di dose misurata. Alcuni software QA permettono agli operatori di radioterapia di inserire anche ulteriori punti di interesse. La differenza di dose puntuale può essere calcolata in modo relativo o assoluto.

4. Materiali e metodi

Per lo svolgimento dello studio sono stati utilizzati dati di pazienti oncologici, diagnosticati con tumore alla prostata e per il cui trattamento è stato scelto un percorso di radioterapia presso l'Istituto IRCC di Candiolo. Nel primo approccio di Machine Learning sono state utilizzate features quantitative caratterizzanti il confronto di dose tra piano di cura e dose effettivamente erogata durante le singole sedute (anche dette frazioni). Nel secondo approccio di Deep Learning si è scelto di confrontare le immagini anatomiche CT (alla base della pianificazione) con le immagini CBCT (acquisite periodicamente nella fase precedente all'erogazione). L'intenzione iniziale dello studio era quella di sviluppare un classificatore che permettesse di individuare tutti i possibili errori riscontrabili in questo distretto. Successivamente, in fase di raccolta dei dati ci si è scontrati con un grosso limite; l'estrema dimensione ridotta del Dataset a disposizione non ha consentito lo sviluppo di un classificatore multiclasse, per questo motivo si è deciso di proseguire lo studio con l'ottimizzazione di classificatori binari per l'individuazione di uno specifico errore: l'errore di preparazione del retto. La scelta è ricaduta su questa tipologia di errore semplicemente perché il più frequente tra i dati raccolti.

4.1 Approccio di Machine Learning

4.1.1 Dataset di riferimento

Per quanto riguarda l'approccio di ML sono stati individuati 20 pazienti per un totale di 152 frazioni. Di queste il 48% appartenente alla classe 0 (no errore di preparazione del retto) e il 52% alla classe 1 (si errore di preparazione del retto).

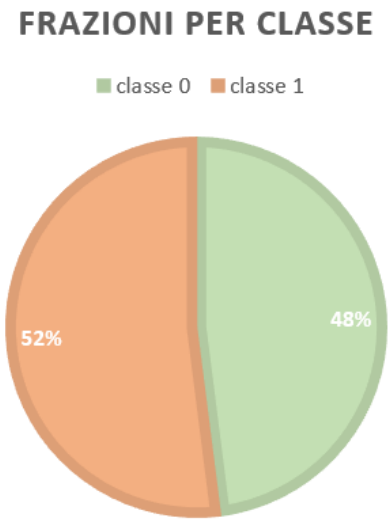

Figura 14 Suddivisione frazioni per classe. 73 elementi su 152 sono di classe 0, 79 su 152 sono di classe 1.

Il dataset totale risulta essere una matrice numerica 152·26. 152 sono le frazioni quindi le osservazioni totali, 25 sono le features caratterizzanti ciascuna osservazione mentre l'ultima colonna riporta la classe di appartenenza di ciascuna osservazione.

Le prime due features sono di tipo informativo e permettono di tenere traccia dell'ID paziente e del numero della frazione. Quelle successive sono differenti metriche, presentate al Capitolo 3, utilizzate per l'analisi della dosimetria in vivo dal software PerFraction:

- Gamma 2D globale sul Beam 1 con soglie 2%-3mm
- Gamma 2D globale sul Beam 1 con soglie 3%-3mm
- Gamma 2D globale sul Beam 1 con soglie 5%-3mm
- Gamma 2D globale sul Beam 1 con soglie 10%-3mm
- Gamma 2D globale sul Beam 2 con soglie 2%-3mm
- Gamma 2D globale sul Beam 2 con soglie 3%-3mm
- Gamma 2D globale sul Beam 2 con soglie 5%-3mm
- Gamma 2D globale sul Beam 2 con soglie 10%-3mm
- Gamma 2D locale sul Beam 1 con soglie 2%-3mm
- Gamma 2D locale sul Beam 1 con soglie 3%-3mm
- Gamma 2D locale sul Beam 1 con soglie 5%-3mm
- Gamma 2D locale sul Beam 1 con soglie 10%-3mm
- Gamma 2D locale sul Beam 2 con soglie 2%-3mm
- Gamma 2D locale sul Beam 2 con soglie 3%-3mm
- Gamma 2D locale sul Beam 2 con soglie 5%-3mm
- Gamma 2D locale sul Beam 2 con soglie 10%-3mm
- Differenza percentuale assoluta tra dose pianificata ed erogata all'isocentro
- DVH sul PTV per il 2% del volume
- DVH sul PTV per il 50% del volume
- DVH sul PTV per il 95% del volume
- DVH sul CTV per il 2% del volume
- DVH sul CTV per il 50% del volume
- DVH sul CTV per il 95% del volume

Tutte le features scelte per la creazione del DataSet sono state concordate con il Fisico Sanitario dell'Istituto di Candiolo. Di seguito si riporta brevemente come sono state ottenute.

GAMMA 2D

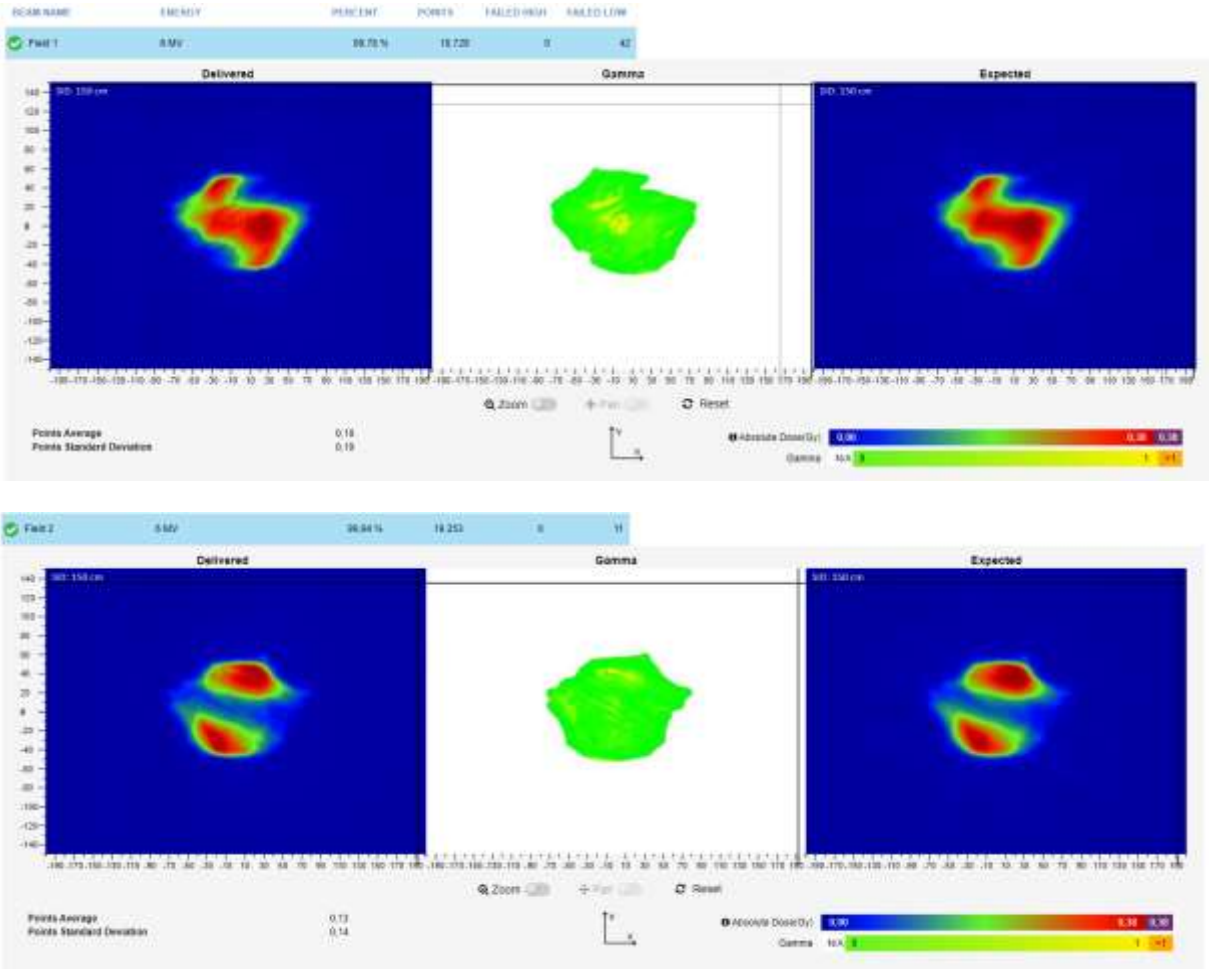

Figura 15 Esempio di Analisi Gamma soddisfacente. Gamma 2D (nel riquadro bianco al centro) tra la dose erogata (nei riquadri blu di sinistra) e la dose pianificata (nei riquadri blu di destra). Nell'immagine in alto il Beam 1 e nell'immagine in basso il Beam 2.

La metrica Gamma è stata utilizzata con diverse soglie, concordate con il fisico Sanitario di Candiolo, per esplorare distribuzioni di dose diverse. La metrica 3%/3mm (alcune volte sostituita dalla più restrittiva 2%/3mm) è la più comunemente usata in campo clinico. Le altre, meno usate portano un'informazione differente, per questo si è scelto di introdurle tutte e verificare in fase di Feature Selection quali fossero più rappresentative. Usando % di scostamento differenti, tra dose pianificata ed erogata, si può discriminare meglio l'entità dell'errore. Infatti, se una percentuale X di punti non supera la soglia più restrittiva 2%/3mm significa che c'è differenza di dose minima ma non si ha veramente misura di quanto questa sia. Se invece gli stessi X punti non superano una % meno restrittiva (10%/3mm) allora l'esito è un altro, c'è stato un importante sovra o sotto dosaggio. Tutte queste metriche sono state calcolate con soglie sia globali che locali per entrambi i fasci (beam1 e beam2).

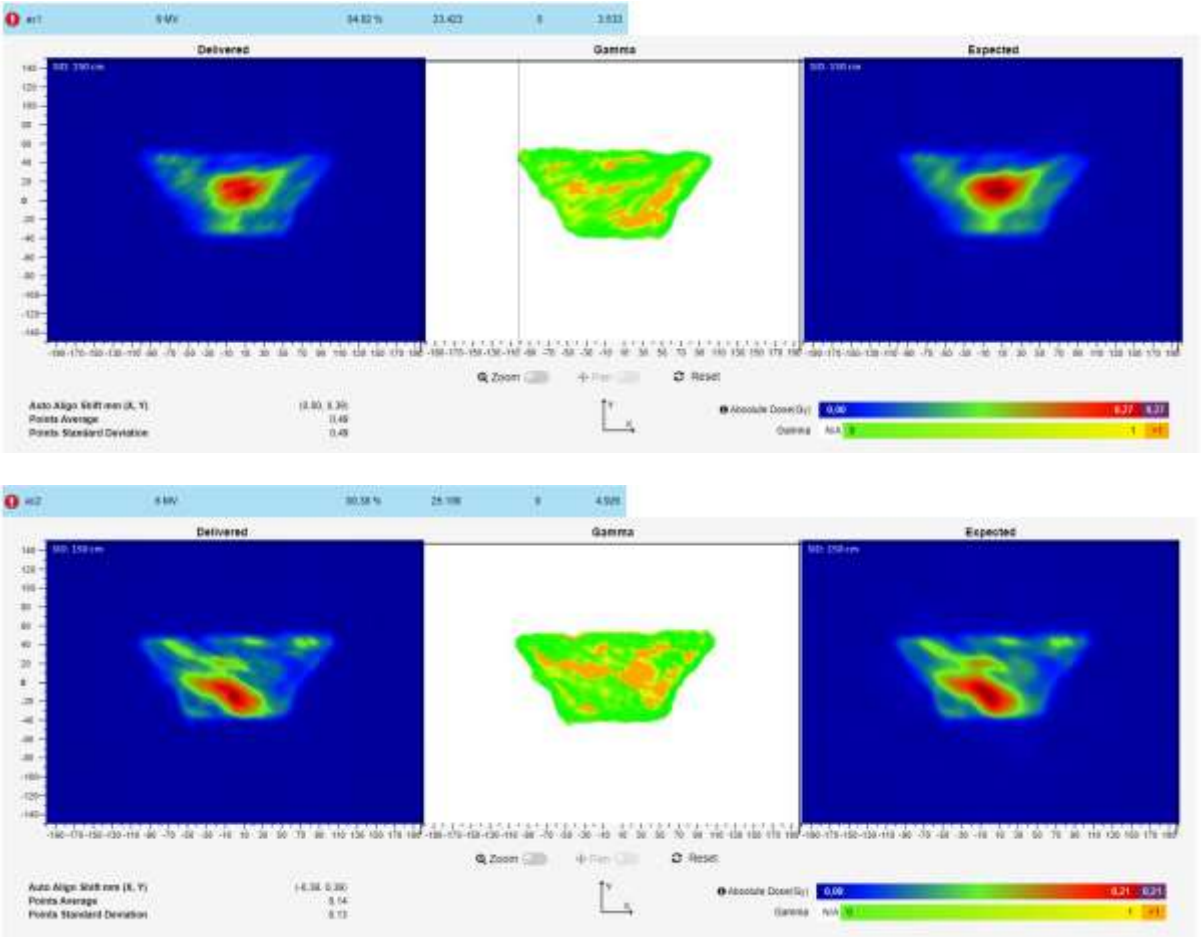

Figura 16 Esempio di Analisi Gamma non soddisfacente, parecchi punti hanno un valore superiore alla soglia 1 di accettabilità. Gamma 2D (nel riquadro bianco al centro) tra la dose erogata (nei riquadri blu di sinistra) e la dose pianificata (nei riquadri blu di destra). Nell'immagine in alto il Beam 1 e nell'immagine in basso il Beam 2.

DOSE ALL'ISOCENTRO

POI NAME	DICOM COORDINATES (mm)			DOSE (Gy)		DOSE DIFF (%/cGy)	
	X	Y	Z	PLANNED	DELIVERED	RELATIVE	ABSOLUTE
act isocentar	-2,32	2,49	-1 443,00	1,951	1,936	-0,7	-1,5

Figura 17 Interfaccia grafica PerFraction. Misura differenziale della dose all'isocentro.

Il valore di differenza relativa di dose all'isocentro è stato semplicemente ricavato dal Software PerFraction.

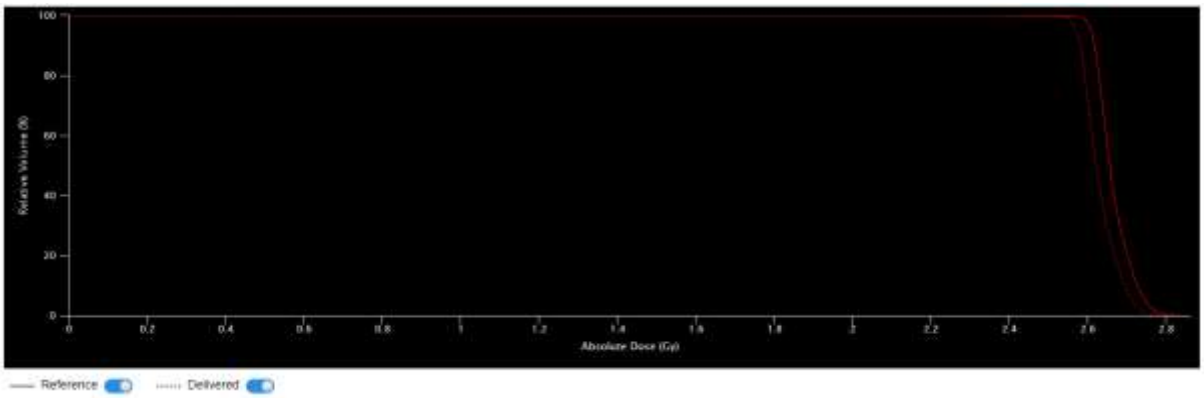

Figura 18 Istogramma cumulativo della distribuzione dose (in ascissa) e volume (% in ordinata) sul PTV.

Con i DVH è possibile tener traccia della copertura di dose su un determinato volume. Si è deciso di considerare come volumi il PTV (planned target volume) e il CTV (clinical target volume). Inoltre, si sono considerati i valori di differenza tra le due curve, pianificata (reference) ed erogata (delivered) per:

- Il 2% del volume: questa metrica consente di verificare la dose massima
- Il 50% del volume: dose media
- Il 95% del volume: dose minima

4.1.2 Assegnazione del ground truth

La classe di appartenenza di ogni frazione è stata assegnata dalla sottoscritta, dopo una fase di preparazione al compito con l'affiancamento del fisico Sanitario dell'Istituto di Candiolo, andando a visualizzare contemporaneamente CT e CBCT per l'individuazione visiva del possibile errore di preparazione del retto.

4.1.3 Divisione DataSet

Siccome è presente una correlazione tra frazioni dello stesso paziente, invece di dividere le osservazioni del DataSet in modo randomico nei Set di dati si è deciso di dividerle tenendo tutte le osservazioni dello stesso paziente nello stesso set di dati. Vista la scarsa numerosità dei dati si è deciso di applicare una procedura di K-fold crossvalidation con k=5 quindi con 4 pazienti per ogni fold. Ad ogni iterazione di allenamento dell'algoritmo 16 pazienti sono usati nel Training set e 4 nel Validation set.

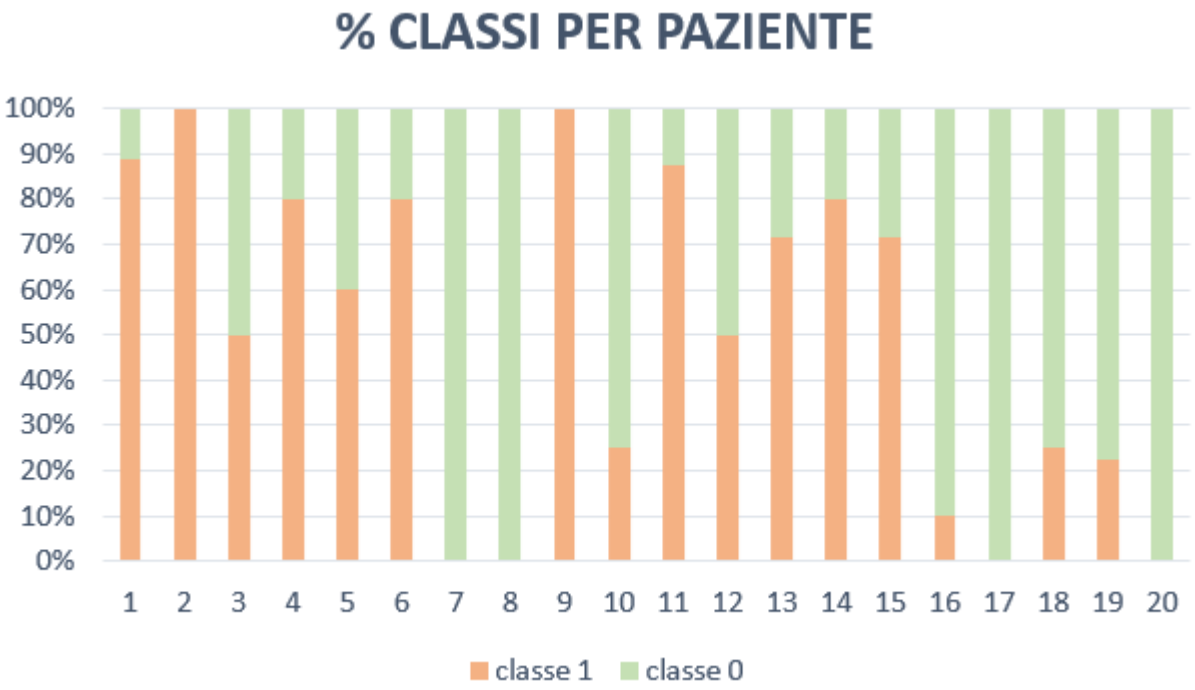

Figura 19 Percentuale di frazioni per ogni classe per ogni paziente.

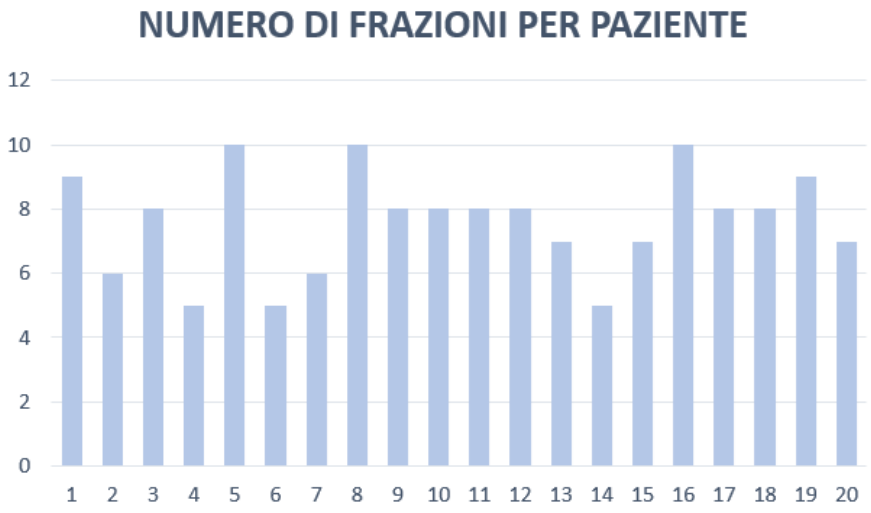

Figura 20 Numero di frazioni per paziente

Figura 21 Suddivisione delle frazioni nei 5 fold per la cross validazione. Diagramma a torta per la rappresentazione percentuale degli elementi delle due classi.

4.1.4 Preprocessing

4.1.4.1 Normalizzazione

Per quanto riguarda la normalizzazione si è scelto un semplice metodo di min max scaling per ogni features, per paziente.

$$g(x) = \frac{(f(x) - f_{min})}{f_{max} - f_{min}}$$

Dove f_{max} è il massimo valore di una certa features tra tutti i valori di features dello stesso paziente e f_{min} il valore minimo.

4.1.4.2 Feature Selection

Come step di preprocessing è stata eseguita Feature Selection (FS) con un metodo wrapper utilizzando un algoritmo genetico (GA), caratterizzato da una strategia di ricerca euristica e randomica per l'individuazione del miglior subset di features.

I GA sono tecniche di ottimizzazione globale ispirate dai meccanismi biologici di selezione naturale e riproduzione. La ricerca probabilistica del miglior subset di features è effettuata usando una popolazione di possibili soluzioni codificate. Una funzione obiettivo (fitness) permette di valutare la bontà di ogni soluzione generata. In questo studio è stata utilizzata una fitness che tenesse in considerazione egualmente le performance di classificazione di entrambe le classi.

$$fitness = 1 - |\frac{sensitibilità + specificità}{2}|$$

4.1.5 Learning

4.1.5.1 SVM

Le Support Vector Machines (SVMs) sono classificatori supervisionati usati in una miriade di applicazioni della vita reale. L'allenamento di un SVM consiste nell'individuare un iperpiano per separare i dati appartenenti a due classi distinte. La posizione di questo iperpiano è solitamente definita con un sottoinsieme di vettori del training set (T) chiamati *'support vectors'* (SVs). Nonostante l'iperpiano consenta una separazione lineare dei dati, gli algoritmi SVMs sono applicabili anche a problemi non lineari grazie alla possibilità di mappare (con kernel) i dati in uno spazio dimensionalmente più grande e nel quale i dati siano linearmente separabili. La selezione di iperparametri come il kernel è uno step computazionalmente caro ma fondamentale per la scelta del giusto modello SVM. Alcune potenziali problematiche individuabili durante l'allenamento di classificatori SVM possono essere legate alla qualità dei dati. Infatti, features o target rumorosi possono impattare negativamente sulle performance del classificatore. Questa problematica è particolarmente visibile in campo medico in cui la maggior parte di test diagnostici non hanno un'accuratezza del 100% e non possono essere considerati un gold standard (ad esempio in fase di classificazione ci possono essere discrepanze tra risultati analizzati da clinici differenti).

Un target rumoroso può avere conseguenze molto pesanti sul comportamento del classificatore quali deterioramento delle performance, aumento di complessità del modello finale rispetto a quanto necessario. Alcuni studi hanno dimostrato che target rumorosi portano ad estrapolare conclusioni incorrette sulle caratteristiche di una popolazione [23].

Vediamo la teoria alla base dei classificatori SVM. Si consideri un set T di t vettori di features per il training $x_i \in \mathbb{R}^D$, i=1,…, t e le corrispondenti classi $y_i \in \{+1,-1\}$ per una classificazione binaria. Vettori di features con label +1 appartengono alla classe positiva C+, gli altri alla classe negativa C-.

SVM lineari

Gli SVM lineari separano i dati nello spazio D-dimensionale tramite un iperpiano definito come:

$$f(x): w^T x + b = 0$$

Dove w è il vettore normale dell'iperpiano, $\frac{b}{||w||}$ è la distanza perpendicolare tra l'iperpiano e l'origine ($b \in \mathbb{R}$). L'iperpiano è posizionato in modo tale che la distanza tra i vettori features delle classi opposte più vicini e l'iperpiano sia massima. Per due classi linearmente separabili i dati del training Set devono soddisfare la condizione:

$$w^T x_i + b \geq +1 \qquad y_i = +1$$

$$w^T x_i + b \leq -1 \qquad y_i = -1$$

Che può anche essere riscritto come:

$$y_i(w^T x_i + b) \geq 0 \qquad y_i \in \{+1, -1\}$$

È un'equazione in cui i vettori features sono posizionati in due iperpiani paralleli con distanza dall'origine rispettivamente $\frac{|1-b|}{||w||}$ e $\frac{|-1-b|}{||w||}$. Non ci sono vettori features tra questi due piani e la distanza tra gli iperpiani è di $\frac{1}{||w||}$. Quindi il massimo margine teorico generabile con l'iperpiano è

$$\varphi(w) = \frac{2}{||w||}.$$

Siccome lo scopo finale dell'SVM è quello di massimizzare il margine di separazione si dovrà cercare di minimizzare il termine $\big|\big|w\big|\big| = \sqrt{w^T w}$.

SVM non lineare

Molti problemi comuni non sono linearmente risolvibili e richiedono una funzione decisionale non lineare. Il Kernel è una funzione matematica utilizzata per trasformare i dati di input in uno spazio dimensionalmente più grande in cui questi siano sperabili. L'introduzione di un kernel permette di

ottenere un iperpiano non lineare. Questo metodo consiste nel definire una funzione (kernel) che calcoli il prodotto scalare di due vettori di features in uno spazio dimensionale non lineare.

Le tipologie più diffuse di kernel sono:

- Lineare
- Polinomiale
- RBF (Radial Basis Function)

In fase di analisi dei dati verranno usati proprio questi tipi di Kernel.

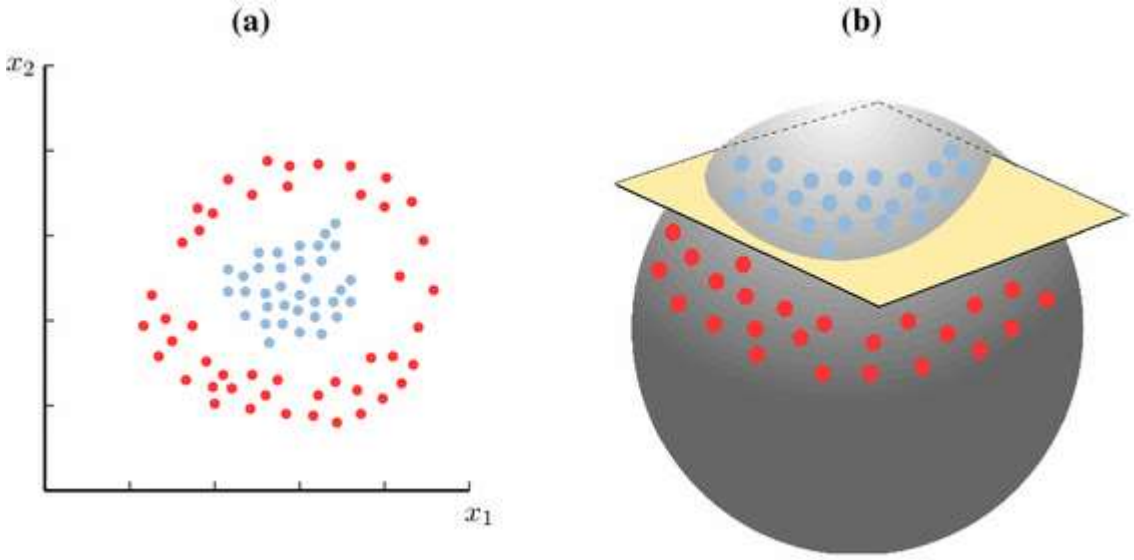

Figura 22 Esempio di dati non linearmente separabili. (a) I punti rossi e azzurri identificano elementi appartenenti a due classi distinte non separabili linearmente. (b) Gli stessi elementi sono mappati in uno spazio dimensionale maggiore dove possono essere separati da un iperpiano (in giallo) [23].

4.1.5.2 NN

Le reti neurali sono modelli di Machine Learning ispirati al meccanismo per cui il cervello umano riesce ad eseguire determinate funzioni. Una rete è composta da unità computazionali (neuroni) collegate tra loro da connessioni (sinapsi) caratterizzate da pesi. Questi pesi vengono ciclicamente modificati per creare un modello di apprendimento dei dati in input.

L'architettura della rete è composta da:

- Unità di input (ricevono i dati di input)
- Unità nascoste (compiono somme pesate dei neuroni di input)
- Unità output (restituiscono un output, una classe ad esempio)

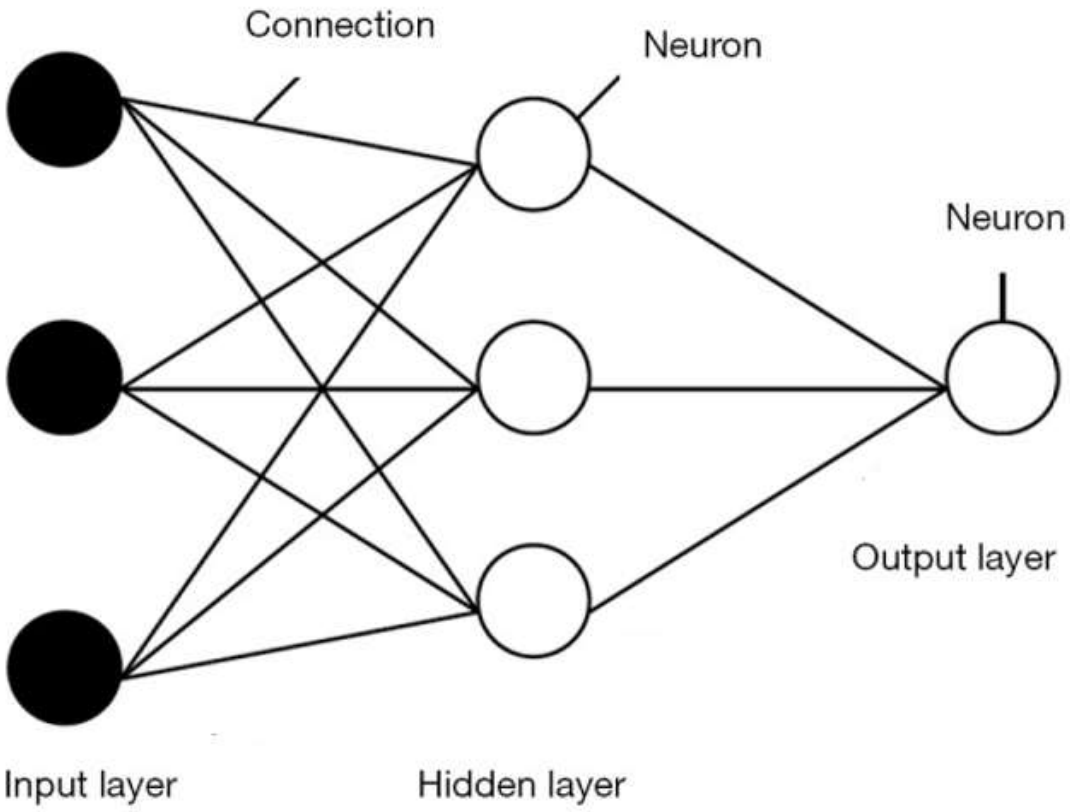

Figura 23 Struttura di una rete feedforward (multilayer perceptron con un layer nascosto e un neurone di output).

Per questo studio si è deciso di utilizzare un 'multilayer perceptron' con un due layer nascosti e un neurone di output.

Gli elementi fondamentali della rete sono le connessioni, con i loro pesi associati, e i neuroni dei layers nascosti e di output che rappresentano le singole unità computazionali. Come mostrato in Figura 24 ogni unità computazionale calcola il risultato dell'applicazione di una funzione di attivazione alla somma pesata degli elementi convergenti nel neurone. In questa tipologia di architettura ogni neurone è connesso a tutti i neuroni del layer precedente e di quello successivo. Il target specifica l'output desiderato per un determinato input. La rete fa propagare il segnale di input da sinistra verso destra e calcola un output sulla base dei pesi attuali (inizializzati randomicamente). Iterativamente poi modifica i pesi in proporzione all'errore calcolato tra l'output desiderato e l'output ottenuto.

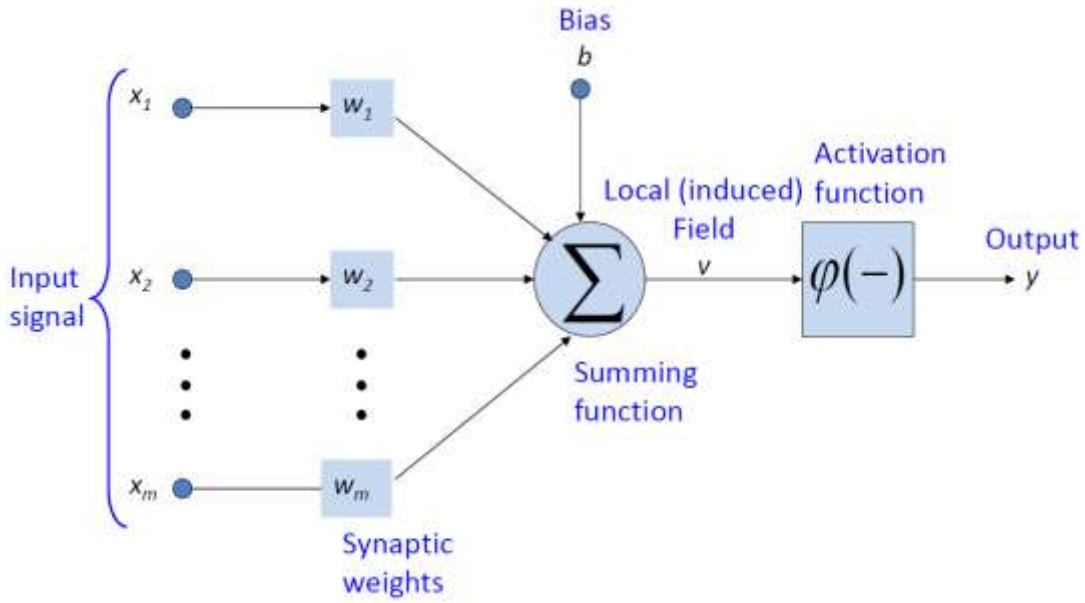

Figura 24 Rappresentazione di neurone, unità fondamentale delle reti neurali.

4.1.6 Evaluation

Nella fase di testing dei due algoritmi di ML sono state eseguite numerose prove combinando strategie diverse di normalizzazione dei dati e di FS.

Per ogni algoritmo:

1. Dati grezzi, considerando tutte le features
2. Dati normalizzati, considerando tutte le features
3. Dati grezzi, dopo aver effettuato FS
4. Dati normalizzati, dopo aver effettuato FS

4.2 Approccio di Deep Learning

4.2.1 Dataset di riferimento

Per quanto riguarda l'approccio di DL sono stati utilizzati i dati di 21 pazienti per un totale di 384 frazioni. Sono state considerate solo le frazioni in cui erano presenti sia la CT di pianificazione (sempre presente), sia la CBCT giornaliera (più difficili da reperire). Nello specifico non è stata presa la CBCT semplice ma si è utilizzata la "CBCT Merged" ovvero la CBCT coregistrata rigidamente alla CT, in fase di posizionamento del paziente sul lettino prima dell'erogazione della terapia, dal software Aria per la gestione dei macchinari.

4.2.2 Preparazione delle immagini

Per la fase di preparazione delle immagini da dare in input all'algoritmo di Deep Learning è stato necessario scaricare dal Database dell'Istituto di Candiolo, tramite il sito SunNuclear, alcune informazioni specifiche sul piano di cura di ogni paziente. Nel dettaglio, sono stati presi il file RT Dose (contenente il piano di dose da erogare al paziente ad ogni seduta) e il file RT structure (contenente l'informazione di contouring delle strutture eseguito dall'oncologo in fase di pianificazione del trattamento radioterapico) entrambi in formato Dicom. Anche l'immagine CT è stata scaricata dal sito SunNuclear mentre la CBCT Merged è stata scaricata direttamente dal Server dell'Istituto di Candiolo tramite query SQL.

Le CT e CBCT Merged, entrambe in formato Dicom, hanno una dimensione di 512·512·N pixels con N variabile da paziente a paziente e una dimensione dei voxel di 1,345·1,345·3 mm. Una volta scaricato tutto il materiale necessario e ordinato per paziente si è proceduto con la compilazione di uno script Matlab per la preparazione delle immagini nel formato finale da dare in input all'algoritmo di Deep Learning.

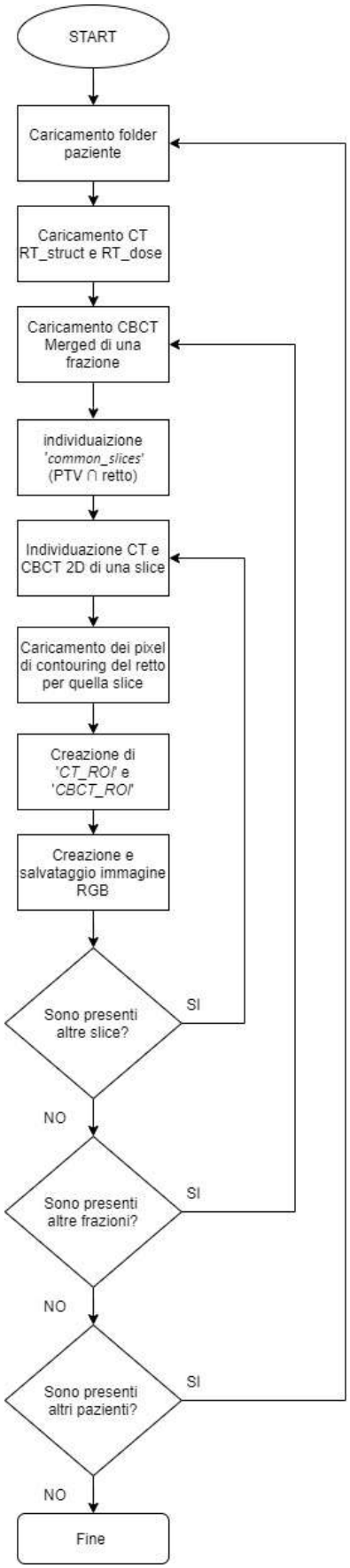

Figura 25 Flow chart del processo di preparazione delle immagini

In Figura 25 è rappresentato il diagramma di flusso del processo eseguito per ottenere le immagini necessarie all'allenamento dell'algoritmo di DL. Il procedimento totalmente automatizzato con uno script Matlab ha permesso di ottenere tutte le immagini necessarie in output ciclando autonomamente su tutte le slice, di tutte le frazioni di tutti i pazienti. Questo è stato reso possibile da una preliminare organizzazione dei dati. Nello specifico si è creata una cartella 'Data' contenente tante sottocartelle identificate tramite l'ID di ogni paziente. All'interno di ogni cartella paziente sono stati inseriti, in cartelle nominate coerentemente per tutti i pazienti, le immagini CT, le immagini CBCT, il piano di dose e il file contenente il contouring delle strutture anatomiche del distretto trattato. Di seguito si riportano in modo più dettagliato gli step rappresentati schematicamente in Figura 25.

Una volta lanciato lo script l'utente ha la possibilità di selezionare la cartella 'Data' in cui sono state ordinate tutte le informazioni necessarie per ogni paziente. A questo punto lo script cicla autonomamente su tutte le cartelle presenti in 'Data' e quindi trattando tutti i pazienti inseriti. Come prima cosa, quando lo script entra nella cartella di un nuovo paziente, si ha il caricamento del file dicom contenente la CT di pianificazione, di quello contenente il piano di dose e quello contenente le informazioni riguardo il contouring delle strutture. Dal file delle strutture lo script individua i VOI (volume of interest) del PTV e del retto, che essendo masse 3D sono caratterizzate dall'unione di tante slice 2D contenenti ROI (region of interest). Si identificano le slice in cui si ha la presenza sia del retto che del PTV e si salva questa informazione nell'array 'common_slices'. A questo punto lo script apre la cartella 'CBCT_Merged' in cui sono presenti, in cartelle separate, tutte le CBCT acquisite durante le diverse sedute di radioterapia. Si apre così un ciclo che permette di analizzare una frazione alla volta. Una volta selezionata la frazione corrente, lo script cicla su tutte le slices presenti nell'array 'common_slices' e per ciascuna si ricava 'ROI_indices', indici dei pixel caratterizzanti il contorno della ROI retto. Dopo aver inizializzato due immagini a zero (=nere) di dimensione 512·512 in una, chiamata 'CT_ROI', va ad inserire in corrispondenza della regione all'interno dei 'ROI_indices' i valori di intensità dei pixel presenti nella stessa regione della CT(precedentemente trasformata in formato uint8) e nell'altra, chiamata 'CBCT_ROI' esegue lo stesso procedimento prendendo i valori di intensità dei pixel presenti sulla CBCT(precedentemente trasformata in formato uint8). Come fase finale di preparazione dell'immagine, nello script viene inizializzata un'immagine RGB nera, matrice 512·512·3. Nella prima dimensione di questa immagine (RGB(:,:,1)) viene inserita l'immagine 'CT_ROI', la seconda dimensione viene lasciata a nero e nella terza (RGB(:,:,3)) viene inserita la 'CBCT_ROI'. Come ultimo step l'immagine RGB viene salvata in formato 'jpg'.

Il procedimento appena spiegato è riferito a una slice, di una frazione di un paziente. Ovviamente lo stesso iter è seguito ciclicamente per tutte le slice di tutte le frazioni di tutti i pazienti fino all'ottenimento del DataSet finale.

4.2.3 Assegnazione del ground truth

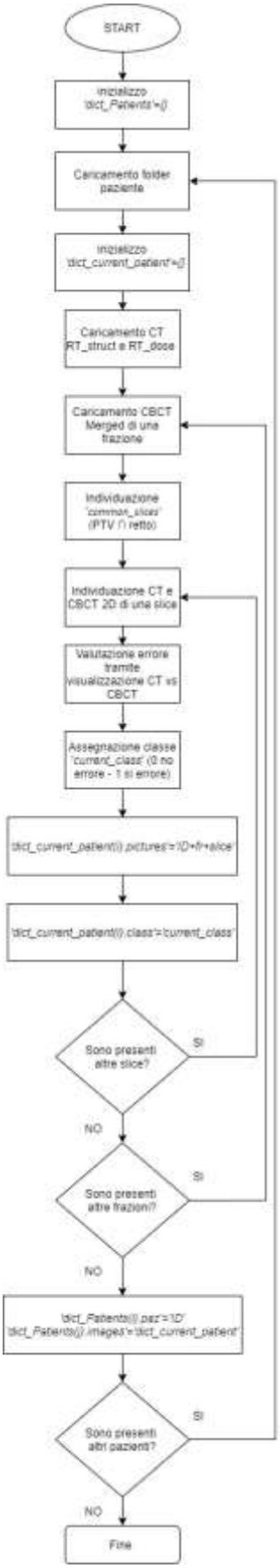

Figura 26 Flow chart del processo per la classificazione delle slices.

A questo punto si è proceduto con l'assegnazione del ground truth. Siccome non è stato possibile far eseguire questa parte ad un operatore esperto, la classificazione delle slice è stata eseguita dalla sottoscritta dopo una preliminare fase di allenamento al compito. Il processo di classificazione è stato eseguito visualizzando una ad una le slices, confrontando CT e CBCT, e assegnando la classe corrispondente: classe 0 corrisponde all'assenza dell'errore mentre classe 1 corrisponde alla presenza dell'errore. Per ovviare parzialmente ai limiti imposti dall'assegnazione del ground truth da parte di un 'rater' non esperto, la classificazione è stata eseguita 3 volte e la classe finale di ogni slice è stata assegnata con tecnica di majority voting.

Vista la necessità di ripetere più volte il processo di classificazione si è optato per implementare uno script Matlab (Figura 26) che velocizzasse il compito attraverso la visualizzazione rapida e in successione di tutte le slice da classificare.

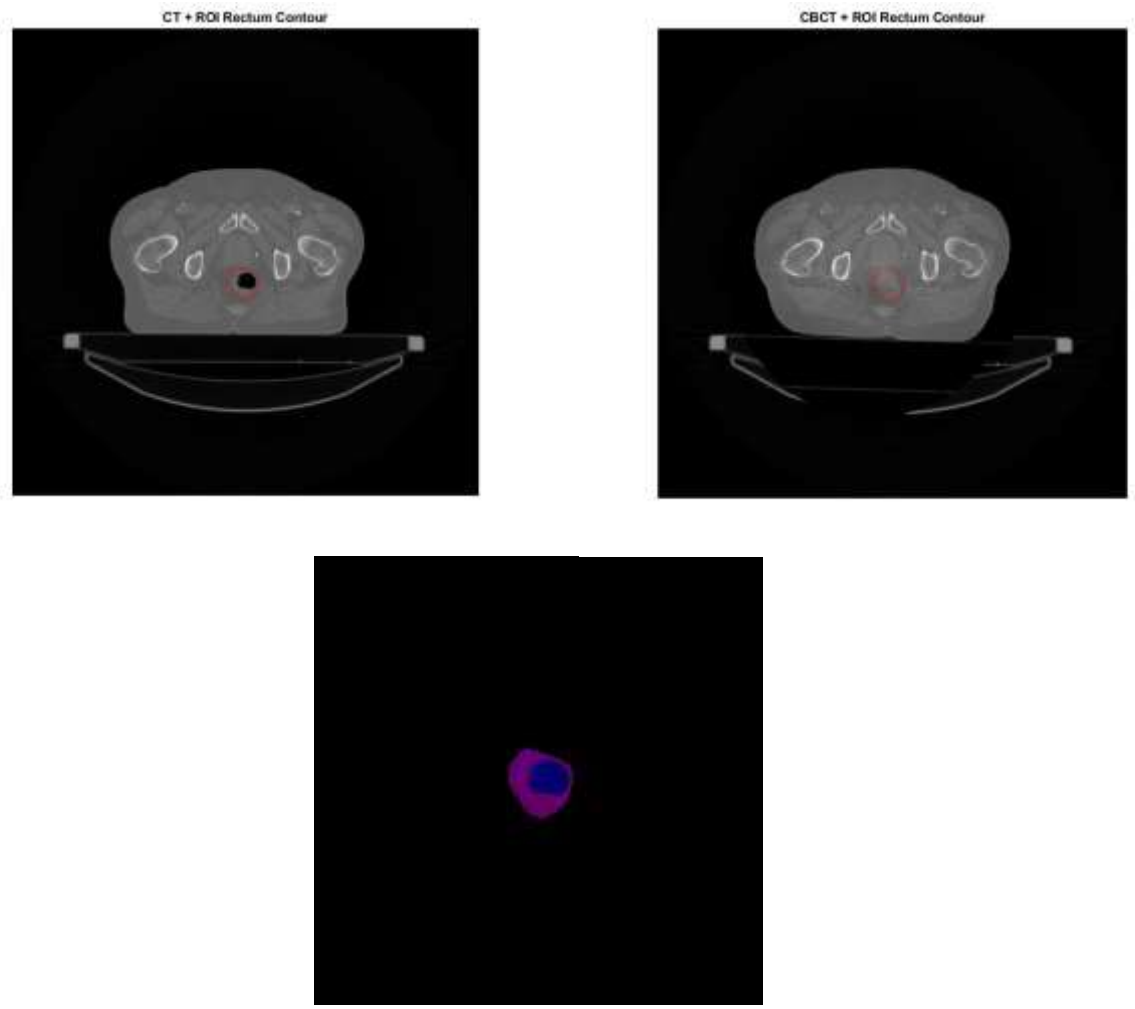

Figura 27 Esempio di errore di preparazione - Slice CT (in alto a sx) e slice CBCT (in alto a dx) con sovrapposto il contouring del retto. Come si nota dalle due immagini nel paziente è presente un visibile errore di preparazione del retto in quanto nella CT è presente una notevole quantità d'aria poi assente nella CBCT. Questa differenza si nota anche nell'immagine RGB (in basso) dove la macchia blu indica una non corrispondenza nell'intensità dei pixel tra le due immagini.

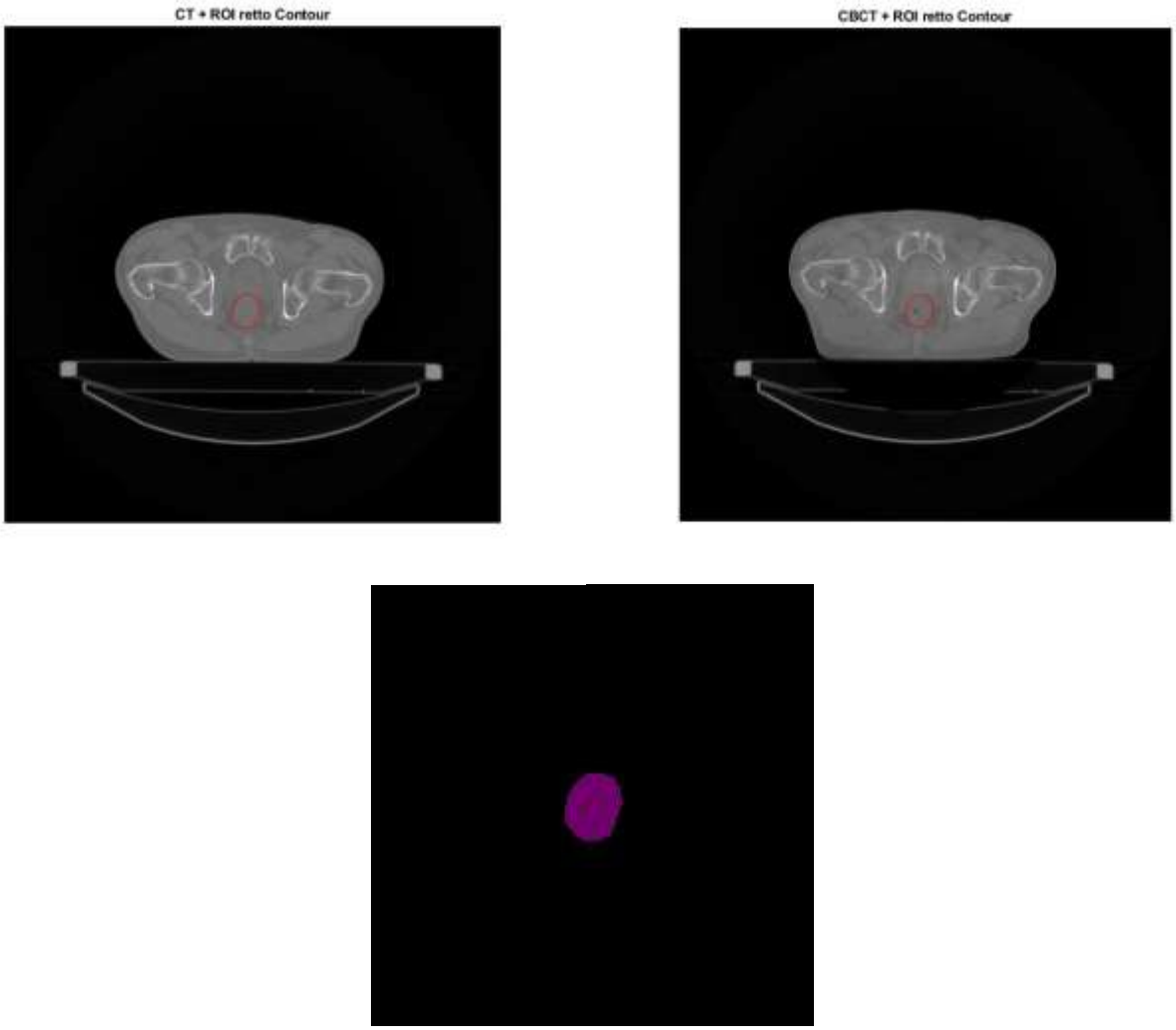

Figura 28 Esempio di errore di preparazione- Slice CT (in alto a sx) e slice CBCT (in alto a dx) con sovrapposto il contouring del retto. Come si nota dalle due immagini nel paziente non è presente un errore di preparazione del retto Questa caratteristica si nota anche nell'immagine RGB (in basso) in cui si nota una omogeneità della ROI.

Per valutare la variabilità intra-operatore, visto che non è stato possibile far svolgere la fase di assegnazione del ground truth da diversi operatori, è stata utilizzata la misura statistica Fleiss' Kappa. Questa misura permette di definire l'affidabilità della concordanza di un certo numero di osservatori quando si assegna una classe ad un certo numero di elementi. La Fleiss' kappa non è altro che la generalizzazione della più famosa Cohen's kappa che funziona solo per le osservazioni di massimo due osservatori. Può essere vista come una valutazione della misura in cui la quantità di accordo tra osservatori eccede quello che ci si aspetterebbe se tutti gli osservatori assegnassero le classi in modo completamente randomico. In sintesi, la kappa fornisce una misura di consistenza delle classi assegnate dai valutatori. Il range di punteggio è tra 0 e 1.

L'affidabilità inter e intra operatore sono misure importanti per definire la validità di un test. In questo caso, per limiti di disponibilità degli operatori è stato possibile valutare solo la variabilità inter operatore. Da questa misura si è ottenuto un valore di kappa pari a 0.917.

4.2.4 Divisione DataSet

In questa fase di preparazione da 21 pazienti e 384 frazioni si sono ottenute 6401 immagini RGB: 5098 di classe 0 e 1303 di classe 1. Non è possibile allenare una rete su un set di immagini pertinenti a un caso specifico e poi testarla su altri dati ma pertinenti sempre allo stesso caso. Allo stesso modo, parlando di immagini 3D, si potrebbe essere tentati di trattare ogni slice come unità indipendente ma sarebbe scorretto. Slice 2D dello stesso volume ma anche slice di frazioni differenti ma dello stesso paziente sono correlate per questo devono essere poste tutte nello stesso set di dati (Training Set, Validation Set o Test set). In caso non venisse fatto le performance sarebbero sovrastimate e non generalizzabili [24].

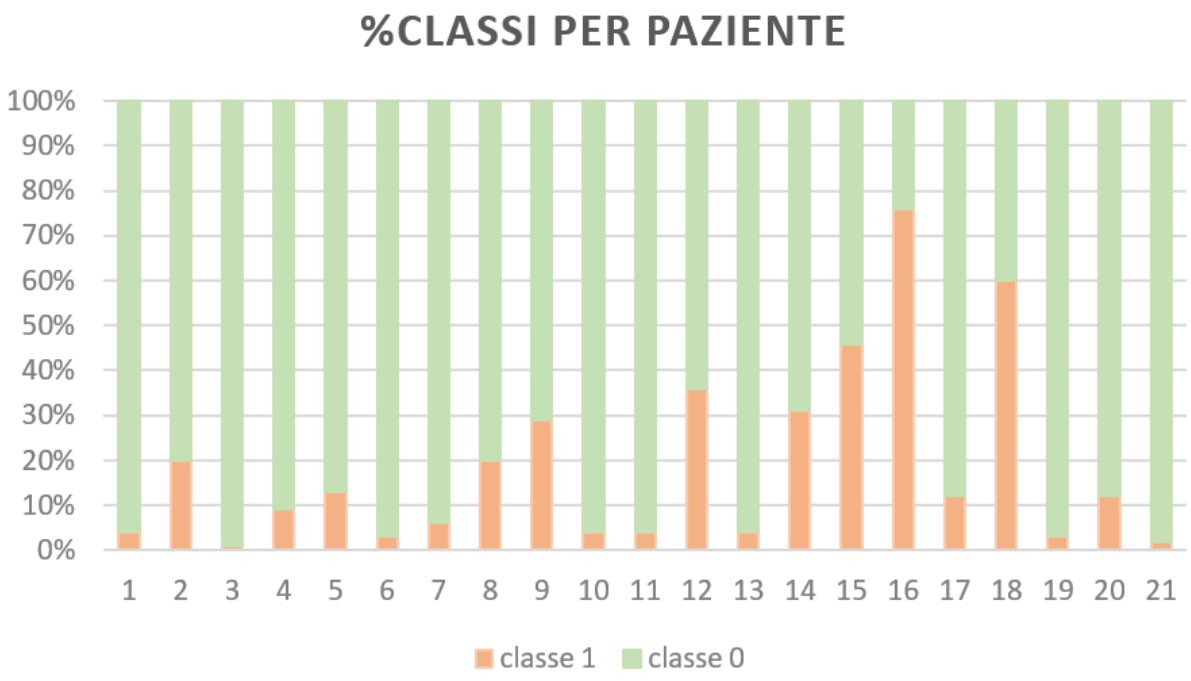

Figura 29 Diagramma a barre. Rappresentazione percentuale delle slices di ogni classe per paziente.

La percentuale di elementi di classe 1 nel 90% dei pazienti (19/21) è inferiore a quella di classe 0. In aggiunta, tra diversi pazienti si notano percentuali molto diverse di slice per ogni classe. Per questo motivo si è prestata molta attenzione nella fase di divisione dei pazienti nei tre Set. Si è deciso di applicare una divisione di dati 70% (15 pazienti) nel Train, 15% (3 pazienti) nel Valid e 15% (3 pazienti) nel Test. La divisione risulta quindi del genere:

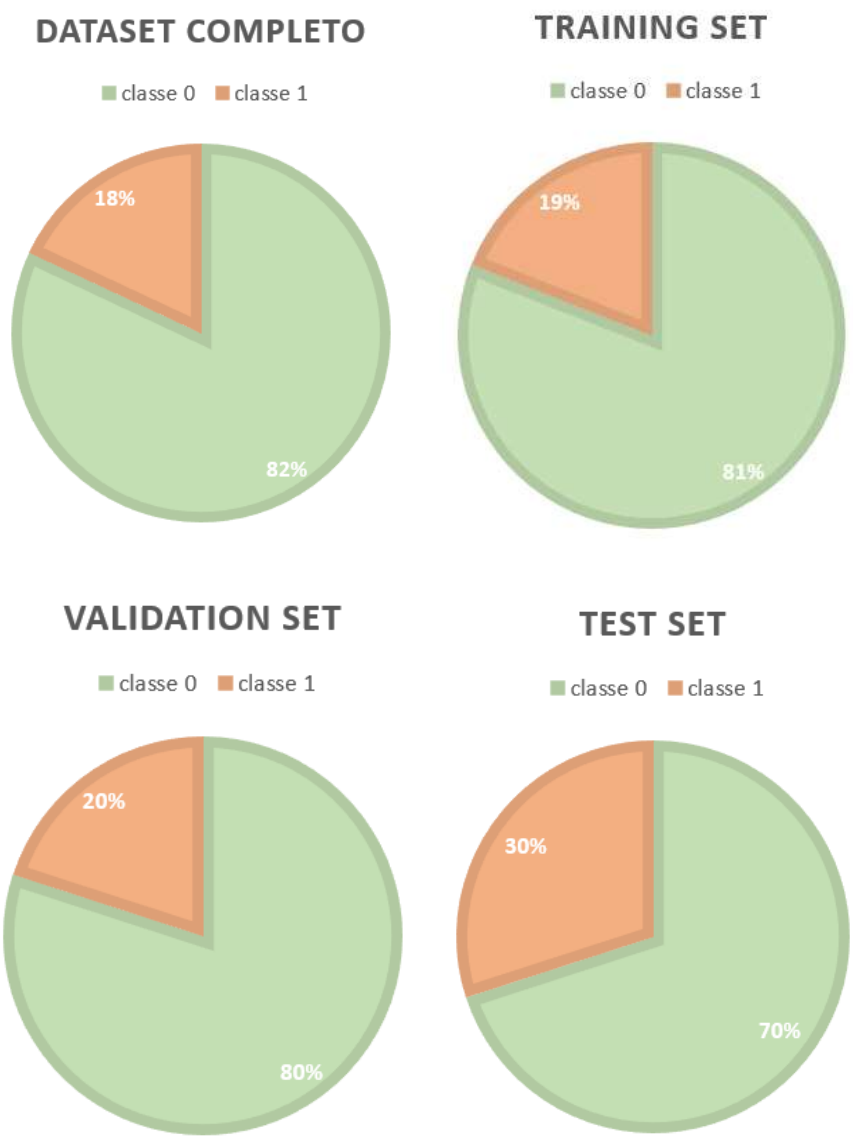

Figura 30 Diagramma a torta. Rappresentazione percentuale della suddivisione delle slice di ogni classe nel Dataset completo, e nei tre set in cui viene diviso: Training Set, Validation Set, Test Set.

4.2.5 Preprocessing

Per quanto riguarda la fase di preprocessing sono stati utilizzate, a scelta, 3 tipologie di normalizzazione: una di min max scaling e due legate al percentile. Il percentile è l'intensità sotto la quale ricade una certa percentuale di pixel. In generale lo scaling α-β percentile permette di riportare la dinamica dei pixel tra α-β ad una dinamica 0-255. Di seguito gli step per il calcolo dei percentili α e β.

- Calcolare l'istogramma cumulativo dell'immagine.

$$C(q) : q \in Q ; C(Q) = \sum_{j=0}^{q} H(j)$$

- Trovare il minor valore in modo che $C(q_a)$ sia maggiore del α% di tutti i pixels.
- Trovare il maggior numero, q_β, in modo che $C(q_\beta)$ sia più piccolo del β% di tutti i pixels.

- Eseguire il mapping lineare secondo la formula

$$g(x,y) = (f(x,y) - f_{\alpha\%}) * \frac{255}{f_{\beta\%} - f_{\alpha\%}}$$

Riportando a 0 i valori di g(x,y)<0 e a 255 i valori di g(x,y)>255.

Le tecniche di normalizzazione usate sono:

- Min max scaling

$$g(x,y) = (f(x,y) - f_{min}) * \frac{255}{f_{max} - f_{min}}$$

- Scaling al 5° e 95° percentile

$$g(x,y) = (f(x,y) - f_{5\%}) * \frac{255}{f_{95\%} - f_{5\%}}$$

- Scaling al 1° e 99° percentile

$$g(x,y) = (f(x,y) - f_{1\%}) * \frac{255}{f_{99\%} - f_{1\%}}$$

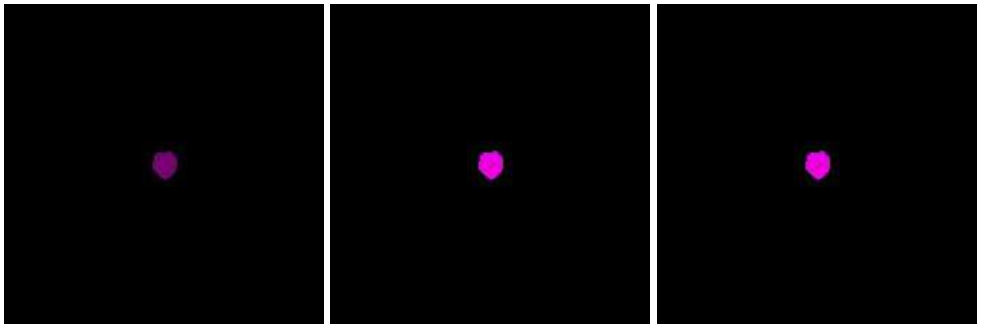

Figura 31 Immagini center cropped, rappresentazione delle diverse tipologie di normalizzazione. Minmax (sx), 5-95 percentile (centro), 1-99 percentile (dx).

In Figura 31 è riportato un confronto visivo delle tre metodiche di normalizzazione applicate sulla medesima slide.

4.2.6 Learning - Convolutional Neural Network

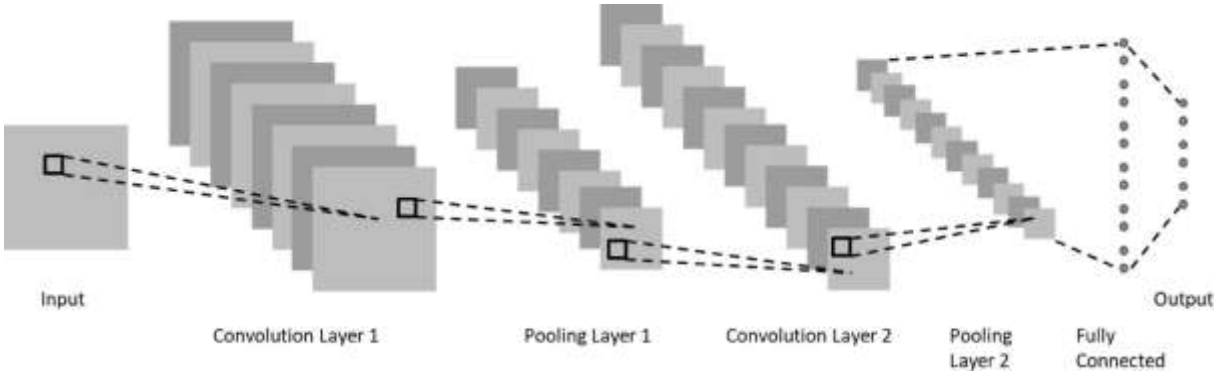

Figura 32 Rappresentazione schematica dell'architettura di una rete neurale convoluzionale [24].

Per la classificazione delle immagini sono state usate le CNN: convolutional neural networks, reti neurali la cui architettura è ispirata al pattern di connessioni dei neuroni della corteccia visiva umana. Neuroni individuali rispondono solo a stimoli ricevuti da ristrette regioni del campo visivo chiamate campi ricettivi. L'analisi visiva generale è resa possibile da un insieme di campi ricettivi sovrapposti che coprono l'intero campo visivo [25].

Le CNN sono composte da una successione di 'blocchi' costitutivi che estraggono le features per permettere di discriminare le immagini date in input in classi di appartenenza differenti. Questi 'blocchi' si dividono in:

- Convolutional layers (CONV) che processano le informazioni di un campo ricettivo
- Correction layers (ReLU), spesso chiamati direttamente 'ReLU' facendo riferimento alla funzione di attivazione (Rectified Linear Unit)
- Pooling layers (POOL) che comprimono l'informazione riducendo la dimensione dell'immagine (spesso sotto campionando).

Questi blocchi costitutivi sono messi in successione, con strutture più o meno lunghe, fino ai layers finali della rete dove avviene la classificazione dell'immagine data in input e il calcolo dell'errore di misclassificazione commesso (differenza tra target e valore predetto). I layers presenti in questa parte terminale della rete sono:

- Fully connected layer (FC)
- Loss layer (LOSS)

Esistono numerose architetture di CNN, ognuna caratterizzata da una diversa successione di CONV, ReLU e POOL layers. In base alla struttura di ogni rete è possibile estrarre informazioni diverse dall'immagine in input e quindi raggiungere obiettivi differenti. Vedremo a breve quali architetture CNN sono più usate per la classificazione di immagini mediche.

4.2.6.1 Transfer learning

Per eseguire un task di classificazione è possibile costruire un modello di architettura 'from scratch' ovvero una CNN personalizzata e su misura per la tipologia di immagini date in input. Gli step principali sono:

- Preparazione dei dati di training, validation e di test
- Costruzione della successione di layers usando librerie specifiche
- Selezione dell'ottimizzatore
- Allenamento la rete
- Test del modello

Solitamente per allenare una CNN 'from scratch' è necessario un Dataset molto grande, caratteristica difficile da trovare soprattutto per i dati in ambito medico. Una soluzione a questo problema è introdotta dal 'Transfer Learning' [26].

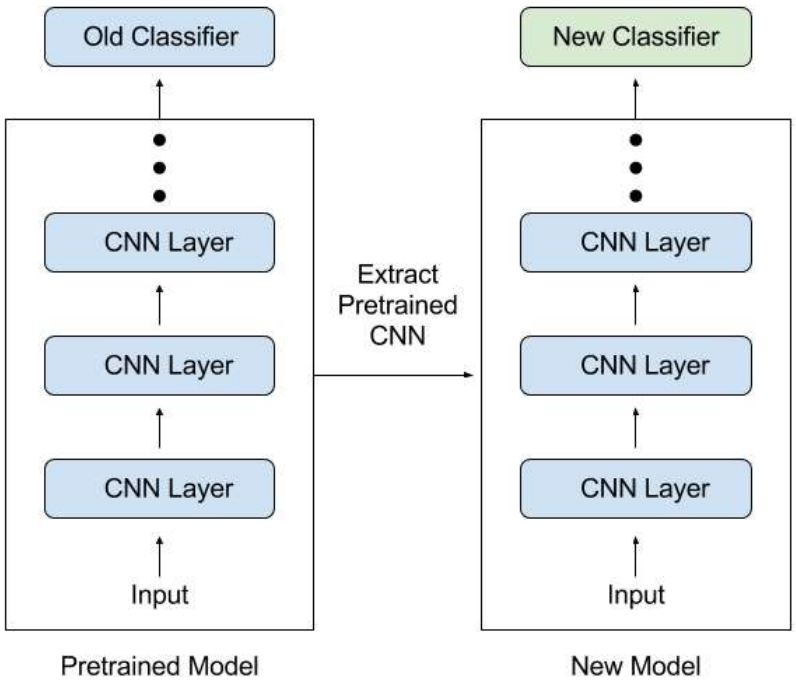

Figura 33 Rappresentazione schematica del concetto di transfer learning [26].

Questa tecnica è molto vantaggiosa perché consiste nel prendere una rete neurale precedentemente addestrata su un problema simile a quello da affrontare, riutilizzando gran parte dei parametri (pesi) già ottimizzati. È quindi sufficiente concentrarsi sull'addestramento degli ultimi layers, quelli più significativi per il risultato finale di classificazione delle features estratte nei layers precedenti [26]. Il transfer learning con immagini mediche solitamente si basa su CNN pre-allenate su immagini naturali (un esempio di dataset comune per pre allenare queste reti di DL è l'ImageNet). Il dataset limitato di immagini mediche è usato per ottimizzare il modello. Durante l'ottimizzazione l'architettura rimane tipicamente fissa e solo un sottoinsieme di pesi viene riallenato [24].

I dataset di immagini mediche sono notoriamente di dimensioni più piccole rispetto ai dataset di immagini naturali, questo perché la raccolta di immagini mediche è un processo dispendioso in termini di tempo e che coinvolge molti step.

Oltre ai vantaggi appena descritti, il 'transfer learning' presenta anche delle criticità. Questa tecnica ovviamente funziona bene solamente se il problema iniziale e quello attuale sono sufficientemente simili. Se la rete è stata allenata su un dataset molto diverso da quello presentato nella seconda fase ovviamente la rete non performerà come ci aspettiamo. Un altro possibile problema è quello dell'overfitting che occorre quando il nuovo modello apprende rumore e dettagli dal training set che impattano negativamente sull'output.

Se si riescono ad evitare questi due problemi principali allora il 'transfer learning' presenta due grandissimi vantaggi: la riduzione delle tempistiche di allenamento, necessità di pochi dati.

4.2.6.2 Crop

L'uso di reti pre-allenate comporta un limite. Le immagini date in input devono avere le stesse dimensioni del layer di input della rete. Tipologie diverse di reti richiedono dimensioni diverse delle immagini. Il dataset in questione, prima di essere dato in pasto alla rete, deve quindi subire una fase di pre-processing. Per questo studio si è scelto di considerarne due diverse tipologie di crop:

- Center crop dell'immagine originale
- Center crop del Bounding Box che circoscriva la ROI del retto + ridimensionamento

Nel primo caso si prende la matrice originale 512·512·3, si individua il centroide della ROI del retto e su questo si centra un quadrato di dimensione N·N (N= dimensione dell'immagine di input richiesta dalla rete scelta). Si esegue così il crop.

Nel secondo caso invece si individua il Bounding Box quadrato (di lato M) più piccolo contenente la ROI del retto e si taglia l'immagine che avrà una generica dimensione M·M (con M<N). A questo punto si

esegue un ridimensionamento dell'immagine per far sì che abbia dimensione N·N. Questa seconda tipologia di crop è stata scelta per diminuire il peso della parte di immagine (messa a nero) che non porta alcuna informazione.

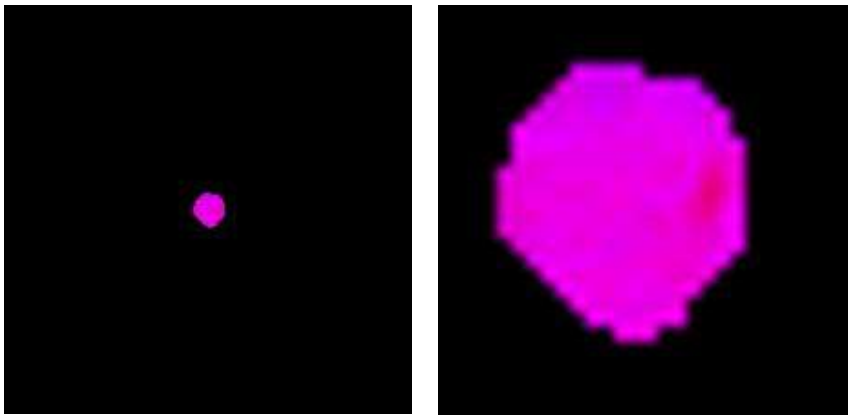

Figura 34 Rappresentazione delle due diverse tipologie di crop dell'immagine. Center crop (a sinistra) e center crop + ridimensionamento (a destra).

4.2.6.3 Dataset augmentation

Gli algoritmi di DL sono in grado di fornire performance decisamente migliori rispetto ad altri metodi di apprendimento ma al costo di richiedere una enorme quantità di dati per l'allenamento. Come detto in precedenza, soprattutto in ambito medico non è quasi mai possibile avere dataset di grandi dimensioni. La tecnica di *data augmentation* è la più semplice per aumentare la numerosità di dati per l'allenamento della rete. Con il data augmentation è possibile generare nuovi dati da quelli esistenti nel dataset, e unirli ai dati originali per aumentare la variabilità del dataset. Le tecniche più comuni adottate per le immagini mediche sono: taglio, traslazione, rotazione e ridimensionamento [24].

Per le prove svolte sono stati impostati i seguenti parametri:

- Riflessione randomica sull'asse x
- Traslazione randomica sull'asse x in un range [-30;30] pixels
- Traslazione randomica sull'asse y in un range [-30;30] pixels
- Scaling randomico sull'asse x in un range [0.9; 1.1]
- Scaling randomico sull'asse y in un range [0.9; 1.1]

Non sono state applicate trasformazioni di shear, riflessioni sull'asse y e rotazioni.

4.2.6.4 Freezing weights + learning rate

Molti studi in cui si è usato il transfer learning hanno optato per ottimizzare il modello di CNN scelto eseguendo ulteriore allenamento su tutti i layers della rete, usando quindi il transfer learning come un metodo di inizializzazione dei pesi.

Con il presupposto che i primi layers delle CNN eseguono operazioni di filtering mentre gli ultimi layers (solitamente fully connected) si focalizzano su features semantiche e di alto livello per obiettivi specifici, altri studi hanno optato per lasciare i primi layers freezati quindi con i pesi della rete pre-allenata andando a ottimizzare solo gli ultimi layers della rete.

Andando a modificare il numero di layers freezati e il learning rate dei diversi layers è possibile modificare i parametri di allenamento (velocità di apprendimento, overfitting ecc..) [24].

4.2.6.5 Class weight

Durante l'allenamento di reti neurali, la 'loss-function' è fondamentale per aggiustare i pesi della rete e creare un modello che si adatti alla tipologia di dati analizzati. Durante la fase di propagazione forward dell'informazione alla rete vengono dati in input i dati del training set, e questa restituisce in output una classificazione per ogni elemento dato in input. Questi risultati vengono confrontati con i valori target e la loss function calcola una penalità per ogni differenza tra target e output della rete. Durante il processo di backpropagation, i pesi allenabili della rete sono aggiustati al fine di raggiungere un modello con minore loss [27].

Nel caso di sbilanciamento delle classi nel Training Set è possibile utilizzare una funzione loss che ne tenga conto. Questo è reso possibile introducendo specifici pesi per ogni classe. In questo caso il peso è associato alla fase di valutazione delle performance; infatti, l'errore di misclassificazione di un elemento della classe meno numerosa peserà di più dell'errore di misclassificazione di un elemento della classe più numerosa. L'idea alla base, quindi, è che non sempre è possibile reperire un numero di elementi uguale per ogni classe, soprattutto in campo medico dove la classe negativa (solitamente associata a patologie, diagnosi ecc) fortunatamente è meno numerosa. Siccome la rete potrà vedere molti meno elementi di una classe si fa in modo che nonostante siano pochi, quando presenti diano un forte impatto sulla modifica dei pesi della rete al fine di migliorare la loro classificazione. Si introduce quindi la weighted – cross – entropy [27]:

$$loss = -\frac{1}{M}\sum_{k=1}^{K}\sum_{m=1}^{M} w_k * y_m^k * \log\left(h_0(x_m, k)\right)$$

Dove:

M= numero di elementi del training set

K= numero di classi

w_k= peso per la classe k

y_m^k = target per il campione m del training per la classe k

x_m= input per il campione m del training

h_ϑ = modello con i pesi ϑ della rete neurale

Nello studio il vettore dei pesi 'classWeights' è stato calcolato in relazione al rapporto di elementi di classe 0 rispetto a quelli di classe 1 nel Training Set. In particolare, è stato ottenuto come:

$$classWeights = \frac{1}{[N0,N1]}$$

$$classWeights = \frac{\text{classWeights}}{\text{mean(classWeights)}}$$

dove $N0$ = numero di elementi del training set appartenenti alla classe 0 (4100 nel nostro caso) e $N1$= numero di elementi del training set appartenenti alla classe 1 (968).

Classe 0	**Classe 1**
0.382	1.618

Dando più peso a un errore di misclassificazione della classe meno numerosa si riesce generalmente a bilanciare l'apprendimento della rete sulle due classi.

4.2.6.6 Google Net

GoogleNet è una deepNetwork a 22 layers nota anche con il nome di Inception V1. Questa rete ha preso ispirazione da LeNet implementando però un modulo iniziale che si basa su piccole convoluzioni permettendo di ridurre notevolmente il numero di parametri (dai 60 milioni di parametri della AlexNet si arriva ai 4 milioni della GoogleNet). Ciò che caratterizza questa rete sono gli *'inception modules'*, 9 in totale, impilati linearmente. Vediamo più nel dettaglio di cosa si tratta.

La necessità di sviluppare la struttura inception è nata dal problema che parti salienti nell'immagine da analizzare possono avere variazioni di dimensione importanti. A causa di queste possibili enormi variazioni nella localizzazione dell'informazione scegliere la giusta dimensione del kernel diventa molto difficile. Infatti, per una distribuzione di informazione più globale si preferisce un kernel più largo mentre per una distribuzione di informazione locale si preferiscono kernel più piccoli. Inoltre, si ricordi che architetture più profonde comportano una maggiore probabilità di overfitting e una maggior difficoltà nell'aggiornamento dei gradienti attraverso l'intera rete. Far susseguire ampie operazioni di convoluzione comporta un elevato costo computazionale. La soluzione è stata quella di provare filtri

con diverse dimensioni ma che operassero allo stesso livello. In questo modo invece di rendere la rete più profonda la si rende semplicemente più ampia. È da qui che nacque '*l'inception module'.* In Figura 35(a) si vede il *'naive inception module'* che permette di eseguire operazioni di convoluzione sull'input con filtri di 3 diverse dimensioni (1x1, 3x3, 5x5). Si nota anche uno step di max pooling. Gli output dopo essere stati concatenati diventano l'input dell''*inception module*' successivo. Per ovviare al problema dell'elevato costo computazionale delle Deep Neural Networks si è pensato di limitare il numero di input aggiungendo una convoluzione 1x1 prima delle convoluzioni 3x3 e 5x5. Queste convoluzioni 1x1 inoltre includono come secondo benefit l'uso di attivazioni ReLU. Il risultato finale è mostrato in Figura 35 (b) [28] [29] [30].

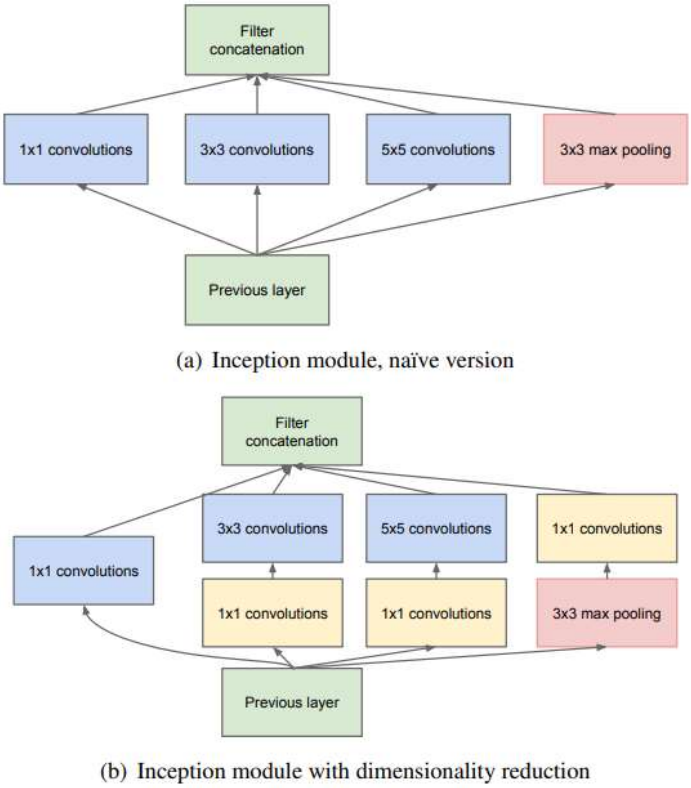

Figura 35 Modulo 'Inception [28][29]'.

In generale, una rete Inception, è una rete composta dai moduli appena spiegati, impilati linearmente.

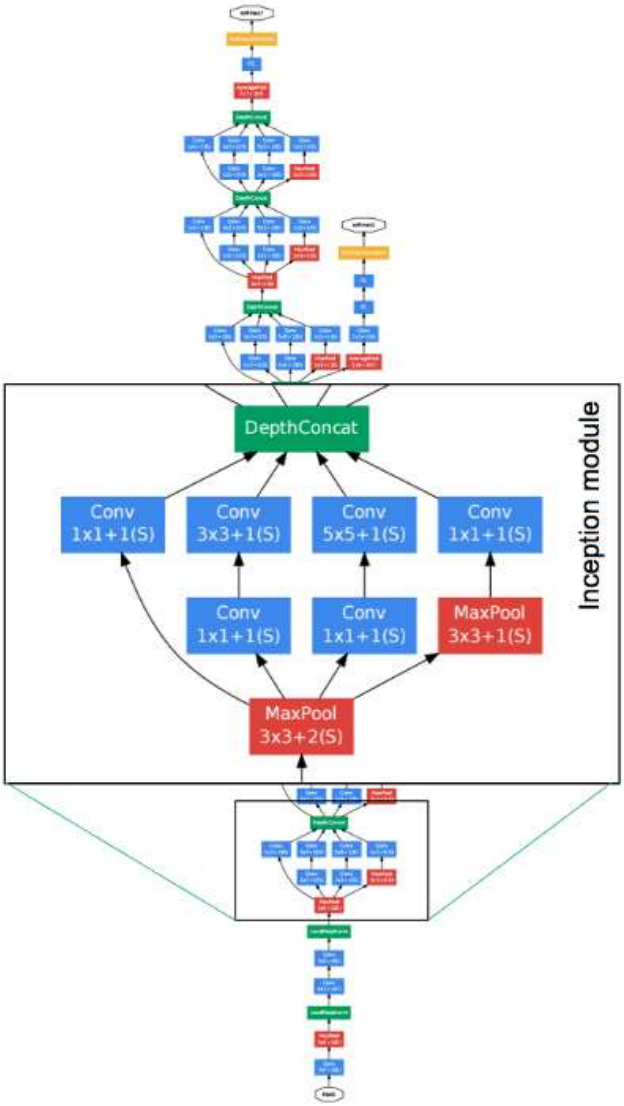

Figura 36 Singolo modulo 'Inception' in una Google Ne t[29].

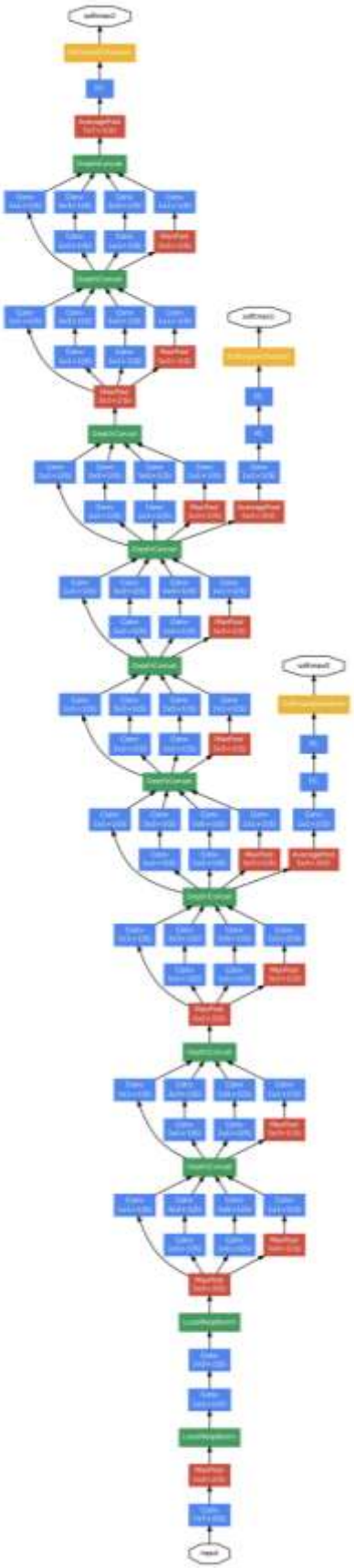

Figura 37 Architettura della Google Net [29].

Vediamo ora la GoogleNet nel suo insieme. La rete riceve in input immagini di dimensione 224·224·3, quindi immagini RGB (con media pari a 0). In tabella X sono ricapitolati poi tutti i layers che concorrono a formare l'architettura GoggleNet.

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Tabella2 Ricapitolazione della struttura inception nella GoogleNet (https://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf)

"#3×3 reduce" e "#5×5 reduce" indicano il numero di filtri 1x1 nel layer di riduzione usati prima delle convoluzioni 3x3 e 5x5.

4.2.6.7 ResNet

Una rete neurale residuale (ResNet) non è altro che una ANN (Artificial Neural network) formata dal susseguirsi di blocchi residuali. Negli ultimi anni le CNN sono state abbondantemente usate per problemi di image recognition, object detection, classificazione di immagini con elevata accuratezza. È stato osservato quanto l'allenamento di reti neurali si complichi con l'aumentare del numero di layers conducendo spesso a due grandi problematiche:

- Vanishing gradient
- Exploding gradient

È qui che sono fondamentali le ResNet con le quali diventa possibile oltrepassare le difficoltà in cui si incorre quando si vuole allenare una rete neurale molto profonda. Per capire il concetto è necessario fare una piccola precisazione. Come noto nelle DNN si utilizza un algoritmo di backpropagation

(assieme ad un metodo di ottimizzazione) per allenare la rete. Calcolando il gradiente della funzione costo apprendiamo di quanto sta sbagliando la rete in fase di predizione e modifichiamo di conseguenza i pesi o parametri della rete. Questa modifica vien fatta retro propagando, layer per layer, l'errore calcolato. Nel caso il numero di layers sia molto alto il gradiente può:

- diventare eccessivamente grande (exploding gradient) generando parametri non gestibili e causando quindi instabilità
- diventare eccessivamente piccolo (vanishing gradient) causando un rallentamento della fase di allenamento del modello.

Il pregio delle ResNet sta nell'utilizzo di una tecnica chiamata 'skip connections', che permette di velocizzare il processo di calcolo del gradiente e aumentare l'efficienza della rete. Grazie a questa skip connection l'output del layers non è più lo stesso. Questa strategia consente di saltare l'allenamento per alcuni layers e di raggiungere direttamente l'output con una connessione differente (mostrata in figura X). Se non usassimo la skip connection l'input 'x' sarebbe moltiplicato per i pesi del layer, sommato a un fattore di bias e il tutto sarebbe sottoposto alla funzione di attivazione f(x). In output avremmo H(x)= f(wx+b) o in breve H(x)= f(x)

Introducendo la skip connection invece l'output cambia in H(x)= f(x)+ x.

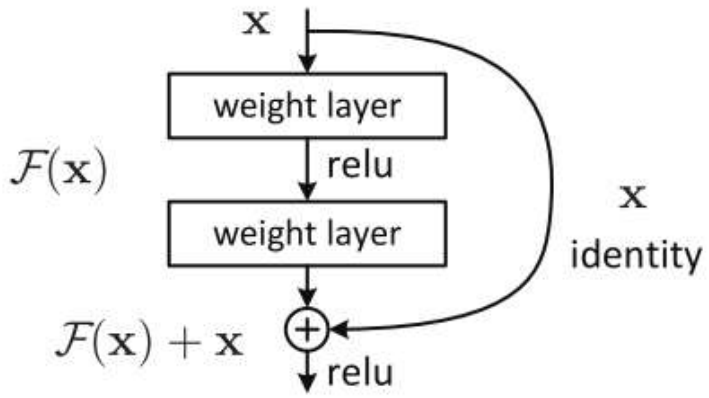

Figure 38 Apprendimento residuale: esempio di skip connection

Esistono diverse architetture che dipendono da come i residuali sono aggregati tra loro. Ad esempio, la ResNet- 50, molto usata, ha 50 layers organizzati in 'residual blocks' [31][32][33][34].

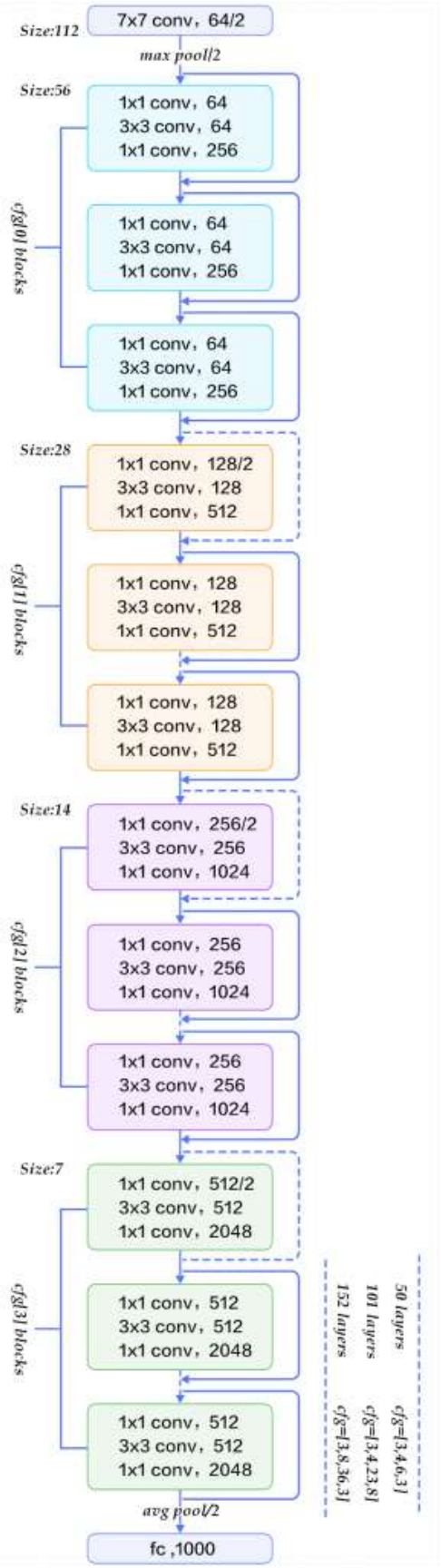

Figura 39 Architettura della rete neurale convoluzionale ResNet.

4.3 Valutazione performance

Le performance di un classificatore binario si valuta solitamente con una confusion matrix. In Figura 40 le colonne rappresentano i valori predetti dal classificatore e le righe sono le classi attuali. TP (True Positive) indica il numero di casi positivi classificati correttamente. FN (False Negative) indica il numero di casi postivi classificati erroneamente come negativi. FP (False Positive) indica il numero di casi negativi classificati erroneamente come positivi. TN (True Negative) indica il numero di casi negative correttamente classificati come tali.

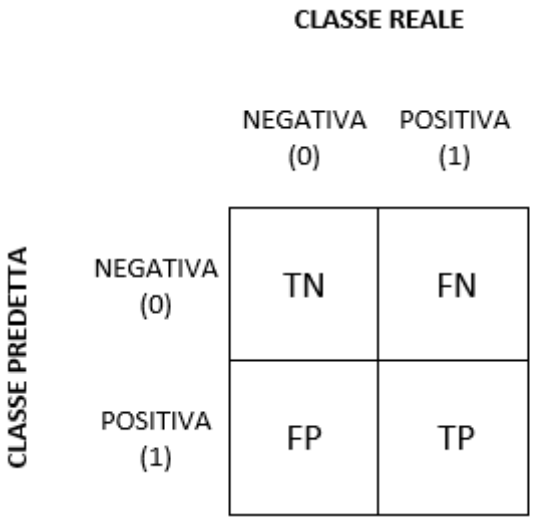

Figura 40 Confusion matrix

La misura più comunemente usata per valutare le performance generali di un classificatore è l'accuratezza. Questa misura però può essere fuorviante quando si è di fronte a uno sbilanciamento del dataset (gli elementi di una classe sono decisamente più numerosi di quelli di un'altra classe). Mentre la sensitività (o recall) misura l'accuratezza di classificazione degli elementi positivi la specificità misura l'accuratezza di quelli negativi. La precision è una misura di correttezza di classificazione da parte del modello allenato. Un alto valore di precision indica un buon classificatore

MISURA	FORMULA
ACCURATEZZA	$\frac{TP + TN}{TP + TN + FP + FN}$
PRECISION	$\frac{TP}{TP + FP}$
SENSITIVITÀ (o RECALL)	$\frac{TP}{TP + FN}$
SPECIFICITÀ	$\frac{TN}{TN + FP}$
G-MEAN	$\sqrt{sensitività \times specificità}$

Tabella 3 Metriche per la valutazione delle performance di classificatori binari

Le prime quattro metriche in tabella X danno informazioni su una classe o sull'altro. Esistono anche metriche che mirano a bilanciare FP e FN. Il G-Mean (geometric Mean) è una metrica che misura il bilanciamento tra le performance di classificazione sia sulla classe più presente, sia su quella meno presente. Un G-Mean basso è indice di performance scadenti per la classificazione della classe positiva anche se gli elementi di classe negativa sono stati classificati correttamente come tali.

$$\mathrm{G} - \mathrm{Mean} = \sqrt{\mathrm{Sensitivity} \times \mathrm{Specificity}}$$

Questa misura è importante per evitare l'overfitting della classe negativa e l'underfitting di quella positiva.

5. Risultati e discussione

5.1 Machine Learning

Quando si parla di ottimizzazione di un classificatore, è necessario trovare un compromesso tra bias e varianza. Il bias è dato dalla differenza tra la predizione del classificatore e la classe reale di un elemento. Modelli con un bias alto semplificano la relazione tra i predittori e la variabile target causando un importante errore di misclassificazione sia sui dati di training che su quelli di test. La varianza invece riflette la variabilità della predizione del modello. Un classificatore con alta varianza si adatta troppo ai dati di training e non riesce a generalizzare nuovi dati in fase di test, di conseguenza in un caso del genere si noteranno buone performance in fase di allenamento ma pessime performance in fase di testing (overfitting).

Il Dataset a disposizione, dopo una lunga fase di ricerca dei dati disponibili è risultato di piccole dimensioni, caratteristica critica per l'allenamento di un buon classificatore. Infatti, quando il Dataset a disposizioni è ridotto diventa difficile estrapolare dai dati per l'allenamento del classificatore un trend che riesca a generalizzare poi nuovi dati, forniti in input al classificatore in fase di test. In questi casi, come riportato da diversi studi, la problematica più facile in cui incorrere è l'overfitting.

Quando si ha a che fare con Dataset di piccole dimensioni gli outliers possono avere un impatto importante sui risultati del classificatore testato. Gli outliers, osservazioni significativamente diverse dalle altre osservazioni, possono essere classificati in naturali e non naturali. I naturali sono dovuti a errori di misurazione, errori nella raccolta dei dati, errori nella trascrizione di valori durante la preparazione del Dataset. Gli outliers naturali invece sono elementi reali ma fuori dalla distribuzione media degli altri valori.

Si è eseguita un'analisi degli outliers tramite boxplot. Da questa prima verifica sono stati individuati due outliers non naturali, causati infatti da un errore di battitura durante la costruzione del Dataset:

- 199 invece di 100
- 9976 invece di 99,76

In questo caso è stato possibile verificare il valore corretto e sostituirlo senza dover rimuovere le due osservazioni. Una volta rimossi questi due outliers sono stati visualizzati nuovamente i boxplot, divisi in due plot distinti, in relazione al range di valori delle features.

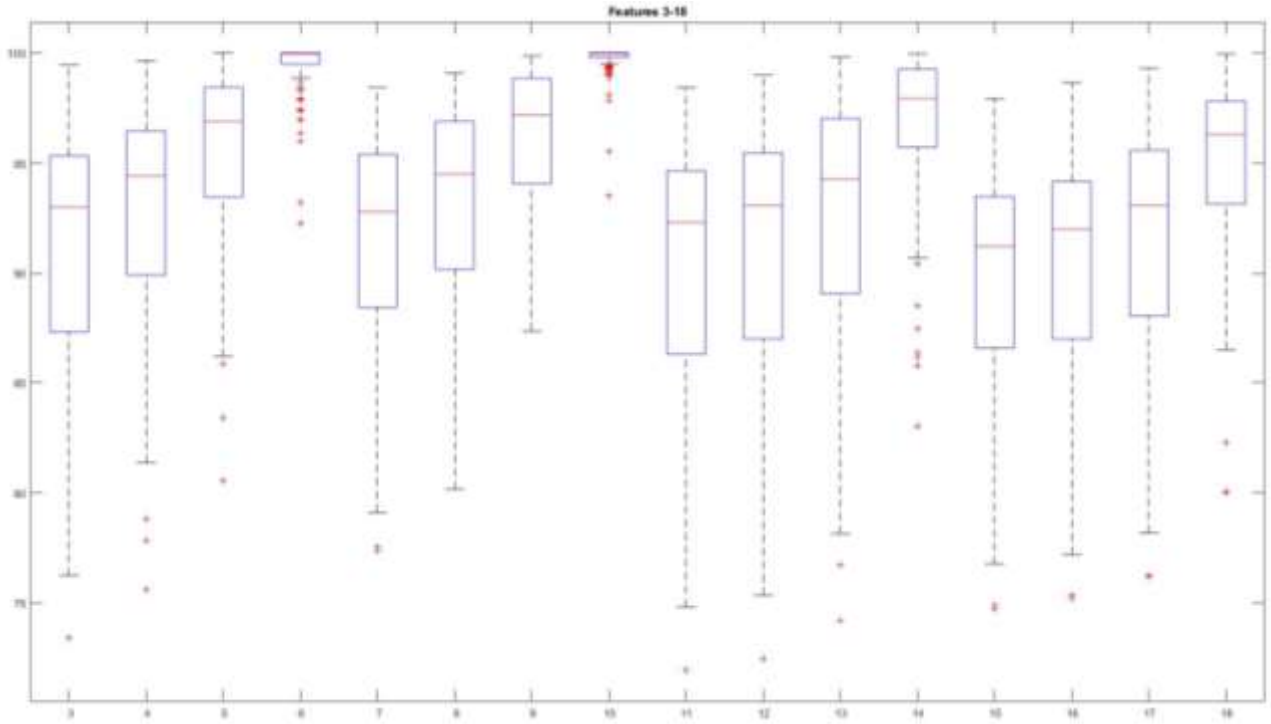

Figura 41 Boxplot degli outliers per le features dalla 3 alla 18

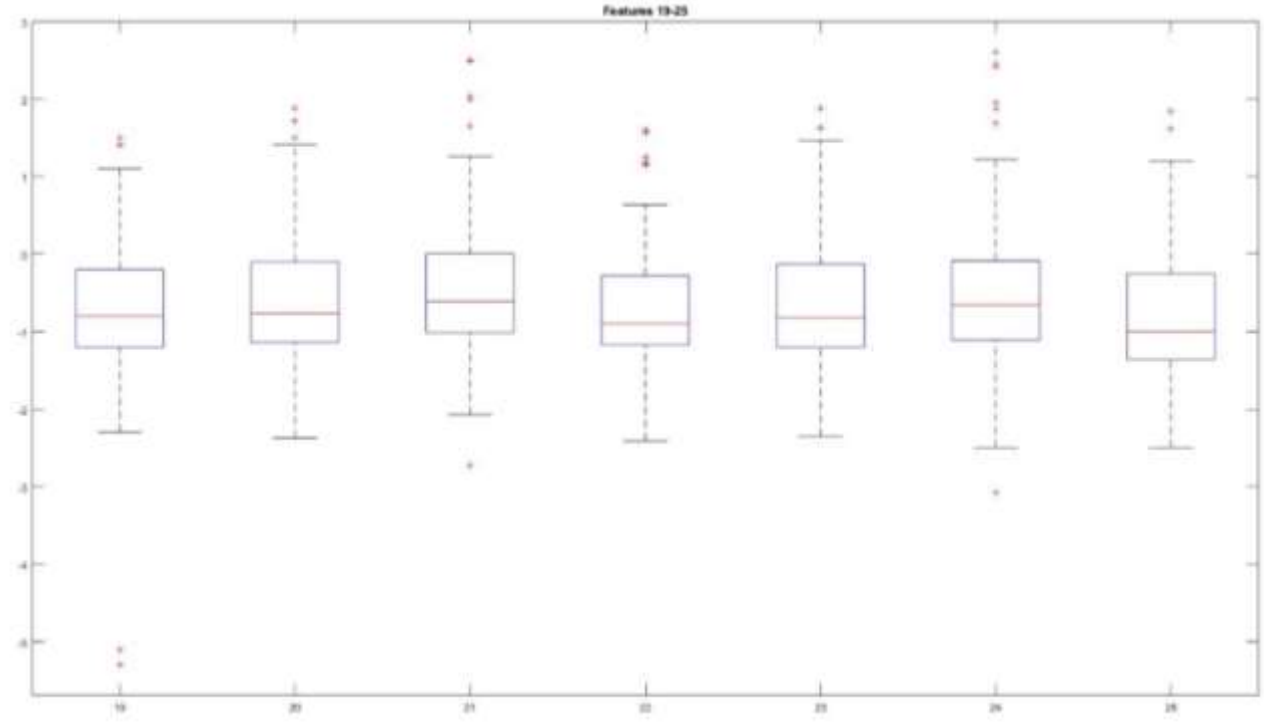

Figura 42 Boxplot per le features dalla 19 alla 25

Come mostrato in Figura 41 e 42 gli outliers risultano essere numerosi. Nel dettaglio è stato individuato un numero di outliers che avrebbe portato alla rimozione di 65 osservazioni su 152, lasciando un

Dataset di sole 87 osservazioni. Per definizione gli outliers sono pochi elementi che si distinguono dall'andamento medio di valori in un'osservazione. Se, come in questo caso, gli outliers sono numerosi (~ 43% del Dataset totale) allora non è corretto escluderli perché portano un'informazione, sono parte integrante dei dati osservati. Questo fenomeno è giustificato anche dalle piccole dimensioni del Dataset. Per questo motivo nessun outliers è stato eliminato. Per quanto riguarda i classificatori si è scelto di testare un SVM e una rete neurale Multilayer Perceptron. Il classificatore SVM è stato testato con 3 kernel differenti (lineare, RBF e polinomiale di secondo grado) utilizzando prima i dati non normalizzati poi quelli normalizzati. I risultati ottenuti sono riportati in tabella 2.

		Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
DATI ORIGINALI	KERNEL LINEARE	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	75,66	accuratezza	51,32
		0	76,00%	24,70%	0	50,70%	48,10%	precision	77,27	precision	53,25
		1	24,00%	75,30%	1	49,30%	51,90%	recall	75,32	recall	51,90
								specificità	76,03	specificità	50,68
	KERNEL RBF	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	100	accuratezza	59,87
		0	100,00%	0,00%	0	42,50%	24,10%	precision	100	precision	58,82
		1	0,00%	100,00%	1	57,50%	75,90%	recall	100	recall	75,95
								specificità	100	specificità	42,47
	KERNEL POLINOMIALE	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	66,94	accuratezza	60,53
		0	68,80%	34,80%	0	68,50%	46,80%	precision	69,36	precision	64,42
		1	31,20%	65,20%	1	31,50%	53,20%	recall	65,19	recall	53,16
								specificità	68,84	specificità	68,49
DATI NORMALIZZATI	KERNEL LINEARE	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	70,89	accuratezza	39,47
		0	70,20%	28,50%	0	34,20%	55,70%	precision	72,20	precision	42,17
		1	29,80%	71,50%	1	65,80%	44,30%	recall	71,52	recall	44,30
								specificità	70,21	specificità	34,25
	KERNEL RBF	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	100	accuratezza	50,66
		0	100,00%	0,00%	0	21,90%	22,80%	precision	100	precision	51,69
		1	0,00%	100,00%	1	78,10%	77,20%	recall	100	recall	77,22
								specificità	100	specificità	21,92
	KERNEL POLINOMIALE	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	99,01	accuratezza	46,05
		0	99,70%	1,60%	0	43,80%	51,90%	precision	99,68	precision	48,10
		1	0,30%	98,40%	1	56,20%	48,10%	recall	98,42	recall	48,10
								specificità	99,66	specificità	43,84

Tabella 4 Risultati delle prove eseguite sul classificatore SVM.

Iniziamo parlando delle prove eseguite sul Dataset non normalizzato. In generale i classificatori con i 3 diversi kernel hanno fornito performance molto varie tra di loro. Infatti, il kernel lineare ha dato risultati medio bassi in fase di training, peggiorati ulteriormente sul Validation Set. L'RBF ha dato overfitting estremo, con accuratezza del 100% sul training e del 59% sul Validation Set. Quello che ha performato meglio, seppur con performance basse (accuratezza del 65%) è stato l'SVM con kernel polinomiale di secondo grado. Questo è l'unico caso in cui il classificatore non è soggetto a overfitting. In questo caso il calo di accuratezza tra Training e Validation è dovuto a un peggioramento di

performance nella classificazione degli elementi di classe 1 mentre le performance sulla classe 0 rimangono invariate. Infatti, complessivamente si notta un valore di specificità uguale mentre si hanno peggioramenti per precision e recall che rispettivamente calano da 69,93% a 64,42% e da 65,19% a 53,16%.

Per quanto riguarda invece le prove eseguite sul Dataset normalizzato, queste hanno portato a performance peggiorative rispetto a quelle ottenute con il Dataset originale, con una forte componente di overfitting per tutti i diversi kernel testati.

Complessivamente, in tutte le prove si è notato un importante calo delle performance tra fase di allenamento e fase di validazione, e quando più attenuato, come nel caso dell'SVM con Kernel polinomiale le performance sono comunque state di bassa qualità, poco più alte di un classificatore 'random guessing'.

Si è quindi deciso di prendere in considerazione un altro algoritmo, per verificare se le performance fossero più legate alla natura dei dati o all'incorrettezza del classificatore scelto. Di seguito si riportano le prove effettuate su una rete neurale multilayer perceptron a due layers nascosti, eseguite come in precedenza sia sui dati grezzi, sia su quelli normalizzati. Le architetture scelte sono: [6 6], [8 8], [10 5], [5 10]. Tenendo in considerazione la componente randomica dei pesi iniziali di queste reti, sono state eseguite 10 ripetizioni per ogni architettura. I risultati, riportati in tabella 3, sono mediati sulle 10 ripetizioni.

		Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
DATI ORIGINALI	[6 6]	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	74,94	accuratezza	57,21
		0	71,50%	21,90%	0	57,00%	42,50%	precision	74,77	precision	59,09
		1	28,50%	78,10%	1	43,00%	57,50%	recall	78,15	recall	57,45
								specificità	71,46	specificità	56,96
	[8 8]	Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	73,63	accuratezza	57,63
		0	76,50%	29,00%	0	64,60%	48,70%	precision	76,54	precision	61,01
		1	23,50%	71,00%	1	35,40%	51,30%	recall	71,02	recall	51,21
								specificità	76,44	specificità	64,58
	[10 5]	Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	74,03	accuratezza	57,86
		0	69,90%	22,20%	0	53,50%	38,20%	precision	73,69	precision	59,03
		1	30,10%	77,80%	1	46,50%	61,80%	recall	77,80	recall	61,84
								specificità	69,94	specificità	53,55
	[5 10]	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	74,83	accuratezza	57,21
		0	71,90%	22,50%	0	51,10%	37,20%	precision	74,93	precision	58,18
		1	28,10%	77,50%	1	48,90%	62,80%	recall	77,50	recall	62,83
								specificità	71,94	specificità	51,13
DATI NORMALIZZATI	[6 6]	Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	68,76	accuratezza	50,96
		0	69,90%	32,30%	0	49,80%	47,90%	precision	70,88	precision	52,86
		1	30,10%	67,70%	1	50,20%	52,10%	recall	67,72	recall	52,06
								specificità	69,89	specificità	49,76
	[8 8]	Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	67,70	accuratezza	49,20
		0	68,50%	33,00%	0	51,20%	52,80%	precision	69,70	precision	51,22
		1	31,50%	67,00%	1	48,80%	47,20%	recall	66,95	recall	47,25
								specificità	68,51	specificità	51,31
	[10 5]	Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	65,62	accuratezza	50,23
		0	60,90%	30,00%	0	45,00%	45,10%	precision	65,94	precision	52,00
		1	39,10%	70,00%	1	55,00%	54,90%	recall	70,01	recall	54,98
								specificità	60,87	specificità	45,08
	[5 10]	Train			Valid			Performance TRAIN(%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	68,78	accuratezza	52,13
		0	69,90%	32,30%	0	49,00%	44,90%	precision	70,90	precision	53,86
		1	30,10%	67,70%	1	51,00%	55,10%	recall	67,72	recall	55,03
								specificità	69,93	specificità	48,99

Tabella 5 Risultati delle prove eseguite sulla rete neurale Multilayer Perceptron.

Anche in questo caso si è notato un peggioramento delle performance sulle prove effettuate con il Dataset normalizzato, basti guardare in tabella 3 la diminuzione di accuratezza su entrambi i set di dati, rispetto alle prove su dai dati originali. Per quanto riguarda invece le prove eseguite sui dati non normalizzati, i risultati sono stati molto costanti tra le diverse architetture testate con un'accuratezza nel Training Set del 74% circa e nel Validation Set del 57% circa. Come anticipato i risultati sulle 4 architetture sono molto simili, la rete scelta come migliore risulta essere quella con due layers nascosti di dimensione 5 e 10. Questa rete, a parità di accuratezza con le altre reti, ha riportato le più alte performance di classificazione della classe 1, con un valore di recall pari a 62,83%. Considerando però la componente randomica data dall'inizializzazione dei pesi, tenuta in considerazione ma non annullabile totalmente, le reti risultano comunque decisamente confrontabili per prestazioni.

Con il confronto delle performance ottenute con un SVM e un Multilayer Perceptron possiamo quindi dire che i risultati deludenti non sono tanto legati al modello scelto per la classificazione quanto più alla natura stessa dei dati. Le problematiche più probabili a questo punto sono:

- Dati insufficienti
- Features non rappresentative del problema

Evitare l'overfitting con un ridotto numero di osservazioni e un numero piuttosto elevato di features può essere complicato. Si è deciso quindi di eseguire uno step di features selection. Siccome la natura dei dati (features numeriche e continue con distribuzione non normale, e relazioni non monotone tra features differenti) non permette di applicare metodi filter basati su misure statistiche, si è scelto di fare FS con un algoritmo genetico (GA). Di seguito si riportano i risultati prima ottenuti con l'SVM poi con il Multilayer.

Con il GA, per l'SVM sono stati testati i 4 kernel più comuni. In tabella 4 si riportano le quattro prove con fitness migliore, prima sui dati originali, poi sui dati normalizzati. E in tabella 5 i risultati delle prove stesse.

DATI ORIGINALI	PROVA 1	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,8	0,2	1000	800	50	1	0,2863	poli2
	PROVA 2	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,9	0,1	1000	800	50	1	0,2721	poli2
	PROVA 3	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,9	0,1	100	80	100	1	0,2774	poli2
	PROVA 4	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,9	0,2	1000	800	50	1	0,2873	poli2
DATI NORMALIZZATI	PROVA 5	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,8	0,2	1000	800	50	1	0,3359	poli2
	PROVA 6	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,9	0,1	1000	800	50	1	0,3253	poli2
	PROVA 7	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,9	0,1	100	80	50	1	0,3390	poli2
	PROVA 8	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	KERNEL
		0,9	0,1	100	80	50	1	0,3369	poli2

Tabella 6 Riepilogo delle migliori prove di FS con GA ottenute: 4 con il Dataset originale e 4 con il Dataset normalizzato. Per ogni prova sono presenti i parametri utilizzati e i risultati di fitness e kernel selezionato dalla prova.

		Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
DATI ORIGINALI	PROVA 1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	61,18	accuratezza	71,05
		0	61,00%	38,60%	0	79,50%	36,70%	precision	62,99	precision	76,92
		1	39,00%	61,40%	1	20,50%	63,30%	recall	61,39	recall	63,29
								specificità	60,96	specificità	79,45
	PROVA 2	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	61,35	accuratezza	72,37
		0	65,80%	42,70%	0	83,60%	38,00%	precision	64,41	precision	80,33
		1	34,20%	57,30%	1	16,40%	62,00%	recall	57,28	recall	62,03
								specificità	65,75	specificità	83,56
	PROVA 3	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	71,71	accuratezza	71,71
		0	78,80%	34,80%	0	86,30%	41,80%	precision	76,87	precision	82,14
		1	21,20%	65,20%	1	13,70%	58,20%	recall	65,19	recall	58,23
								specificità	78,77	specificità	86,30
	PROVA 4	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	65,79	accuratezza	71,05
		0	74,00%	41,80%	0	76,70%	34,20%	precision	70,77	precision	75,36
		1	26,00%	58,20%	1	23,30%	65,80%	recall	58,23	recall	65,82
								specificità	73,97	specificità	76,71
DATI NORMALIZZATI	PROVA 5	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	74,67	accuratezza	65,79
		0	85,60%	35,40%	0	82,20%	49,40%	precision	82,93	precision	75,47
		1	14,10%	64,60%	1	17,80%	50,60%	recall	64,56	recall	50,63
								specificità	85,62	specificità	82,19
	PROVA 6	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	76,97	accuratezza	67,11
		0	82,50%	28,20%	0	76,70%	41,80%	precision	81,65	precision	73,02
		1	17,50%	71,80%	1	23,30%	58,20%	recall	71,84	recall	58,23
								specificità	82,53	specificità	76,71
	PROVA 7	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	79,28	accuratezza	65,79
		0	83,20%	24,40%	0	74,00%	41,80%	precision	82,99	precision	70,77
		1	16,80%	75,60%	1	26,00%	58,20%	recall	75,63	recall	58,23
								specificità	83,22	specificità	73,97
	PROVA 8	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	79,61	accuratezza	65,79
		0	85,30%	25,60%	0	79,50%	46,80%	precision	84,53	precision	73,68
		1	14,70%	74,40%	1	20,50%	53,20%	recall	74,37	recall	53,16
								specificità	85,27	specificità	79,45

Tabella 7 Performance delle prove di FS su GA con classificatore SVM per Dataset originale e Dataset normalizzato.

Come già visto prima della FS, dalla tabella 4 si nota nuovamente come il kernel polinomiale di secondo grado predomini sulle altre tipologie di kernel testate, consentendo performance decisamente migliori. In questo caso tra prove su dati normalizzati e prove su dati non normalizzati si nota un andamento molto diverso. Con l'utilizzo del Dataset normalizzato si presenta nuovamente il fenomeno dell'overfitting, seppur accompagnato da un'accuratezza leggermente più alta. Con l'utilizzo del Dataset non normalizzato invece si nota ancora una volta come un Dataset così ridotto non sia rappresentativo del problema nella sua totalità. Infatti, le prove con il Dataset originale riportano performance migliori in fase di validazione piuttosto che in fase di allenamento. Quando il dataset è limitato si fa la cross-validazione proprio per non far dipendere le performance dai dati che ricadono

nel Validation Set. In questo caso, nonostante la cross-validazione, non si riesce a evitare questa problematica. Probabilmente le immagini di validazione sono molto simili, in termini di features, ad alcune immagini del Training Set in cui il classificatore funziona bene.

Per quanto riguarda i dati non normalizzati la prova con fitness migliore (pari a 0,2721) ha riportato la scelta di kernel polinomiale, 14 features selezionate e accuratezza del 61,35% sul Training Set e del 72,37% sul Validation Set. Le performance migliori riguardano la classificazione degli elementi di classe 0 rispetto a quelli di classe 1.

Per quanto riguarda invece i dati normalizzati la prova con fitness migliore (pari a 0,3253) ha riportato la scelta del kernel polinomiale, solo 5 features selezionate e accuratezza del 76,97% sul Training Set e del 67,11% sul Validation Set. Anche in questo caso gli elementi di classe 0 hanno un maggior numero di corretti classificati rispetto agli elementi di classe 1.

Complessivamente il numero di features totali selezionate è molto variabile. Non si riscontra concordanza neanche tra le tipologie di features scelte nelle varie prove di FS.

Di seguito, in tabella 6 e 7, sono riportate le prove di FS eseguite tramite algoritmo genetico con il Multilayer Perceptron come modello per l'ottimizzazione.

DATI ORIGINALI	PROVA 1	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,8	0,2	1000	800	50	1	0,2451	[5 10]
	PROVA 2	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,8	0,2	1000	800	50	1	0,2588	[5 10]
	PROVA 3	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,9	0,1	1000	800	50	1	0,2536	[8 8]
	PROVA 4	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,8	0,2	1000	800	100	2	[0,2689]	[10 5]
		0,8	0,2	1000	800	100	2	[0,2689]	[8 8]
DATI NORMALIZZATI	PROVA 5	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,8	0,2	1000	800	50	1	0,3100	[5 10]
	PROVA 6	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,8	0,2	1000	800	50	1	0,2994	[5 10]
	PROVA 7	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,9	0,1	1000	800	50	1	0,3100	[10 5]
	PROVA 8	PC	PM	NIND	NGENITORI	NITER	NRIP	FITBEST	RETE
		0,8	0,2	1000	800	100	1	0,3126	[8 8]

Tabella 8 Riepilogo delle migliori prove di FS con Multilayer Perceptron ottenute: 4 con il Dataset originale e 4 con il Dataset normalizzato. Per ogni prova sono presenti i parametri utilizzati e i risultati di fitness e architettura della rete selezionato dalla prova.

		Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
DATI ORIGINALI	PROVA 1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	71,20	accuratezza	57,67
		0	72,50%	30,00%	0	62,30%	46,70%	precision	73,37	precision	60,53
		1	27,50%	70,00%	1	37,70%	53,30%	recall	69,98	recall	53,33
								specificità	72,52	specificità	62,37
	PROVA 2	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	75,59	accuratezza	59,62
		0	69,80%	19,00%	0	59,40%	39,50%	precision	74,37	precision	61,57
		1	30,20%	81,00%	1	40,60%	60,50%	recall	80,94	recall	60,44
								specificità	69,81	specificità	59,35
	PROVA 3	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	74,99	accuratezza	60,23
		0	72,00%	22,20%	0	54,70%	34,70%	precision	75,02	precision	60,94
		1	28,00%	77,80%	1	45,30%	65,30%	recall	77,78	recall	65,36
								specificità	71,97	specificità	54,67
	PROVA 4.1	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	74,87	accuratezza	56,35
		0	67,90%	18,70%	0	53,20%	40,70%	precision	73,29	precision	57,81
		1	32,10%	81,30%	1	46,80%	59,30%	recall	81,27	recall	59,26
								specificità	67,94	specificità	53,20
	PROVA 4.2	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	71,18	accuratezza	59,87
		0	67,70%	25,60%	0	56,60%	37,10%	precision	71,37	precision	61,05
		1	32,30%	74,40%	1	43,40%	62,90%	recall	74,38	recall	62,92
								specificità	67,71	specificità	56,56
DATI NORMALIZZATI	PROVA 5	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	63,84	accuratezza	51,77
		0	57,60%	30,40%	0	43,50%	40,60%	precision	63,99	precision	53,23
		1	42,40%	69,60%	1	56,50%	59,40%	recall	69,58	recall	59,41
								specificità	57,62	specificità	50,84
	PROVA 6	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	65,68	accuratezza	52,51
		0	65,20%	33,80%	0	48,50%	43,90%	precision	67,27	precision	54,15
		1	34,80%	66,20%	1	51,50%	56,10%	recall	66,14	recall	56,16
								specificità	65,17	specificità	48,54
	PROVA 7	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	52,72	accuratezza	66,15
		0	64,60%	32,40%	0	50,40%	45,10%	precision	54,47	precision	67,37
		1	35,40%	67,60%	1	49,60%	54,90%	recall	54,91	recall	67,62
								specificità	50,34	specificità	64,56
	PROVA 8	Train			Valid			Performance TRAIN (%)		Performance VALID (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	69,99	accuratezza	52,85
		0	70,40%	30,40%	0	55,50%	49,50%	precision	71,81	precision	55,06
		1	29,60%	69,60%	1	44,50%	50,50%	recall	69,59	recall	50,47
								specificità	70,42	specificità	55,42

Tabella 9 Performance delle prove di FS su GA con classificatore Multilayer Perceptron per Dataset originale e Dataset normalizzato

Da una prima analisi delle prove di FS eseguite con Multilayer Perceptron si nota come siano ancora presenti alcune caratteristiche riscontrate con le prove eseguite prima della FS. Ad esempio, in Tabella 6 tra le architetture di rete selezionate in fase di FS si vedono comparire 3 architetture sulle 4 testate. In precedenza, infatti avevamo notato come le performance tra i multilayers con diverse architetture fossero del tutto confrontabili, rispetto alle prove sull'SVM in cui il kernel polinomiale ha performato molto meglio rispetto agli altri. Successivamente, da un'analisi più approfondita si nota come, rispetto

alla FS su SVM, in questo caso il fenomeno di overfitting è presente sia sui dati originali che sui dati normalizzati, anche se i primi restituiscono risultati leggermente più soddisfacenti con un'accuratezza delle diverse prove compresa tra il 57% e il 60%. Questi risultati ovviamente non possono essere considerati sufficienti per l'ottimizzazione di un buon classificatore.

Complessivamente nessun classificatore, né con dati normalizzati né con dati originali, con o senza FS ha restituito risultati minimi accettabili.

Un'ultima soluzione, per la gestione di DataSet piccoli è quella di combinare diversi classificatori per ottenere un risultato più completo e attendibile. Nel nostro caso questa opzione è stata esclusa viste le performance di partenza dei singoli classificatori. Andando ad unire le classificazioni di due modelli pressoché randomici non si traggono benefici rilevanti. In sintesi, i risultati ottenuti dal classificatore con dati quantitativi (metriche per il confronto dei valori di dose in vivo con quelli panificati) non hanno riportato performance soddisfacenti.

Questo risultato non ci stupisce per diverse ragioni. Innanzitutto, il Dataset di partenza, come spesso accade per studi in ambito medico, è di dimensioni molto ridotte. Va ricordato anche che i valori di differenza di dose utilizzati come osservazioni per l'ottimizzazione di questo classificatore non sono legati univocamente all'errore di preparazione. Possono infatti coesistere diversi errori, più o meno importanti che sommandosi portano ad una determinata distribuzione di dose o a un determinato scostamento tra dose pianificata e dose erogata. Ovviamente lo sviluppo di un classificatore non binario, sulla carta sicuramente più adatto alla risoluzione del problema, è stato escluso fin da subito causa la dimensione del Dataset.

Viste le performance scadenti, ma giustificate, in fase di Training non si è ritenuto opportuno proseguire con una fase di Testing. Si è deciso però di intraprendere una strada diversa per la classificazione di questo errore: un'analisi di immagini con tecniche di Deep Learning.

Questa scelta è facilmente giustificabile infatti oggigiorno la prima fase di classificazione degli errori possibilmente presenti durante l'erogazione della terapia viene fatta dal tecnico di radioterapia tramite la visualizzazione e il confronto di immagini CT (di pianificazione) e CBCT (eseguite qualche istante prima dell'erogazione della terapia). Metodologia seguita anche dalla sottoscritta per l'assegnazione della classe nelle osservazioni del Dataset finora utilizzato. Fallito il tentativo innovativo di classificare gli errori con dati dosimetrici si è deciso di provare a intraprendere una classificazione con reti neurali per l'analisi di immagini, in particolare con le reti neurali convoluzionali (CNN) [35].

5.2 Deep Learning

Come modelli per la classificazione di immagini sono state scelte due reti neurali convoluzionali (CNN). In particolare, si è deciso di testare la GoogleNet e la ResNet.

Per ogni rete sono state effettuate 6 prove combinando 3 tipologie di normalizzazione e 2 tipologie di crop dell'immagine.

1. **Normalizzazione**: Min-Max scaling
   Ridimensionamento: center crop dell'immagine intera
2. **Normalizzazione**: 5-95 percentile
   Ridimensionamento: center crop dell'immagine intera
3. **Normalizzazione**: 1-99 percentile
   Ridimensionamento: center crop dell'immagine intera
4. **Normalizzazione**: min-max scaling
   Ridimensionamento: center crop della ROI del retto + resize
5. **Normalizzazione**: 5-95 percentile
   Ridimensionamento: center crop della ROI del retto + resize
6. **Normalizzazione**:1-99 percentile
   Ridimensionamento: center crop della ROI del retto + resize

5.2.1 GoogleNet

Tutte le 6 prove sono state svolte con gli stessi parametri di rete:

- 100 layers freezzati (è stata utilizzata la tecnica di transfer learning, introdotta al Capitolo 5)
- Learning rate dell 'ultimo layer (FullyConnectedLayer) =10
- Learning rate dei layers precedenti = 0,001
- 10 epoche di allenamento

		Train			Valid			Test			Performance TRAIN (%)		Performance VALID (%)		Performance TEST (%)	
Center crop	min max	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	94,87	accuratezza	70,75	accuratezza	89,74
		0	95,80%	9,10%	0	64,60%	4,50%	0	94,60%	21,40%	precision	83,65	precision	40,25	precision	86,34
		1	4,20%	90,90%	1	35,40%	95,50%	1	5,40%	78,60%	specificità	95,80	specificità	64,55	specificità	94,59
											Gmean	93,32	Gmean	78,53	Gmean	86,23
	5-95 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	95,22	accuratezza	94,18	accuratezza	87,18
		0	95,30%	5,10%	0	93,50%	3,00%	0	87,00%	12,40%	precision	82,64	precision	78,79	precision	74,58
		1	4,70%	94,90%	1	6,50%	97,00%	1	13,00%	87,60%	specificità	95,29	specificità	93,47	specificità	87,01
											Gmean	95,12	Gmean	95,23	Gmean	87,29
	1-99 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	94,59	accuratezza	95,52	accuratezza	87,63
		0	94,50%	5,00%	0	95,50%	4,50%	0	87,70%	12,40%	precision	80,28	precision	84,21	precision	75,54
		1	5,50%	95,00%	1	4,50%	95,50%	1	12,30%	87,60%	specificità	94,49	specificità	95,52	specificità	87,66
											Gmean	94,76	Gmean	95,52	Gmean	87,61
Center crop + resize	min max	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	96,33	accuratezza	76,57	accuratezza	84,46
		0	97,00%	6,40%	0	71,50%	3,00%	0	84,20%	14,90%	precision	87,96	precision	45,94	precision	70,08
		1	3,00%	93,60%	1	28,50%	97,00%	1	15,80%	85,10%	specificità	96,98	specificità	71,46	specificità	84,20
											Gmean	95,27	Gmean	83,26	Gmean	84,64
	5-95 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	95,88	accuratezza	97,01	accuratezza	86,12
		0	95,80%	3,70%	0	97,60%	5,20%	0	83,30%	7,50%	precision	84,34	precision	90,71	precision	70,72
		1	4,20%	96,30%	1	2,40%	94,80%	1	16,70%	92,50%	specificità	95,78	specificità	95,57	specificità	83,33
											Gmean	96,03	Gmean	96,17	Gmean	87,81
	1-99 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	96,15	accuratezza	95,82	accuratezza	84,92
		0	96,10%	3,80%	0	97,90%	12,70%	0	82,70%	10,00%	precision	85,49	precision	91,41	precision	69,35
		1	3,90%	96,20%	1	2,10%	87,30%	1	17,30%	90,00%	specificità	96,15	specificità	97,95	specificità	82,68
											Gmean	96,16	Gmean	92,48	Gmean	86,29

Tabella 10 Performance delle prove eseguite sulla rete GoogleNet.

Complessivamente le prove (le cui performance sono riassunte in tabella 8) hanno riportato risultati soddisfacenti, sia su Training Set e Validation Set che sul Test Set con un'accuratezza di classificazione tra il 96% e l'84%. Le prove con taglio center crop hanno complessivamente performato leggermente meglio rispetto alle prove con center crop e ridimensionamento. Le prove con center crop e normalizzazione con metodica percentile hanno riportato performance pressoché identiche, è stata scelta come migliore la rete con normalizzazione 1-99 percentile per uno 0,7% in più di corretti classificati nella classe 0, differenza irrilevante. In figura 43 è mostrato l'andamento di accuratezza e loss function durante il processo di allenamento della rete individuata per migliori performance. Da questa immagine si nota un buon andamento decrescente della loss function nella fase iniziale (sia per Training che per Validation Set), seguito poi da un andamento pressoché costante durante le epoche successive, stabilizzato intorno ad un valore di ~0,1. Per quanto riguarda il grafico dell'accuratezza invece, si ha una buona crescita (sia per Training che per Validation Set) durante la prima epoca e poi una stabilizzazione per le epoche successive nella zona tra il 90% e il 100%.

Per quanto riguarda i risultati riportati nelle confusion matrix di tabella 8, si ha un buon bilanciamento dei corretti classificati per le due classi e valori complessivi di accuratezza tra il 94% e il 96% sul Training, tra il 70% e il 97% sul Validation e tra l'84% e l'89% sul Test. I risultati più scarsi di accuratezza sul Validation sono dati dalle prove eseguite sul Dataset normalizzato con tecnica min-max scaling,

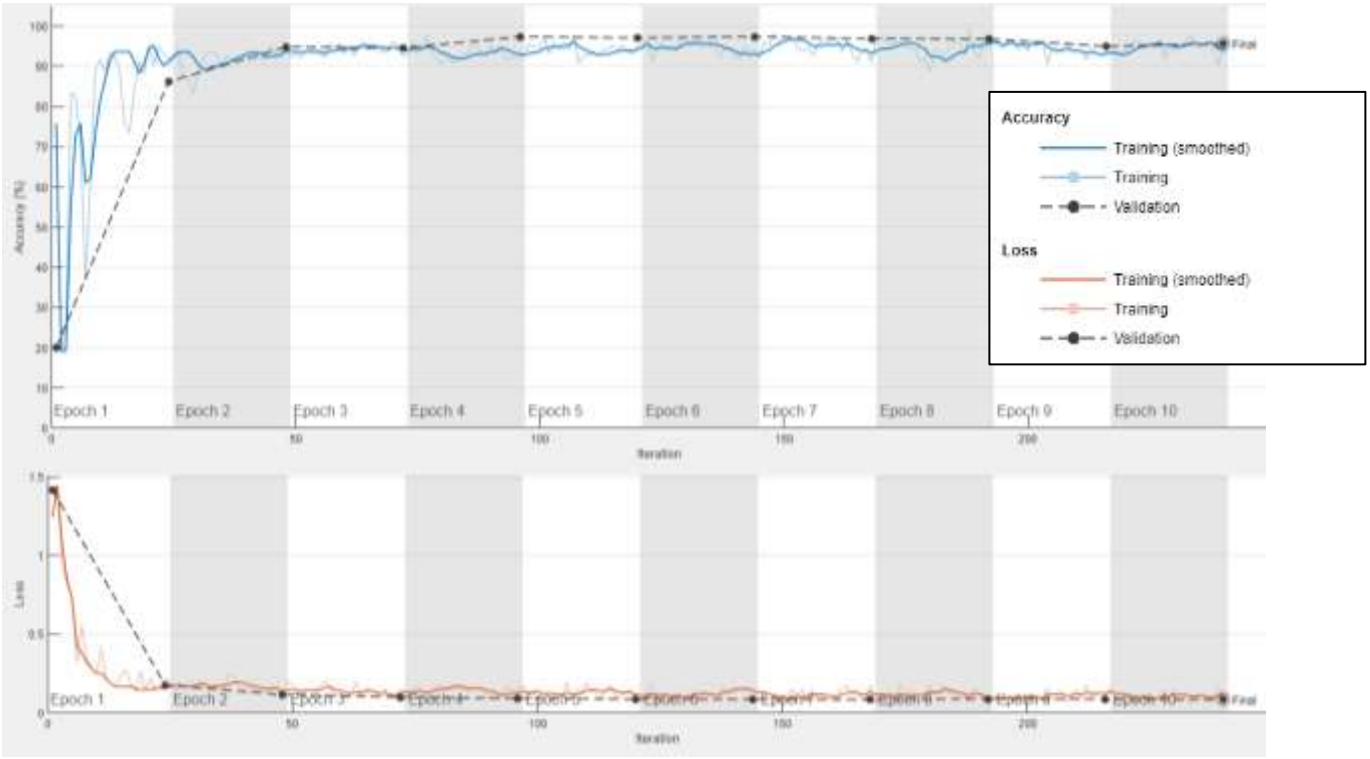

Figura 43 Grafico dell'allenamento della GoogleNet allenata sul dataset normalizzato con 1-99 percentile e immagini ottenute tecnica center crop.

Individuato il Set di immagini migliore per proseguire (normalizzate tra il 1° e il 99° percentile e ottenute con metodo center crop), sono state eseguite ancora alcune prove riducendo i parametri allenabili, rischiando un leggero underfitting, ma con l'obiettivo di diminuire il gap di performance tra Training/Validation Set e Test Set. In particolare, il learning rate del layer finale è stato aumentato nella prima prova (tabella 9) da 10 a 50 e nella seconda prova da 10 a 100. Di seguito i risultati ottenuti.

		Train			Valid			Test			Performance TRAIN (%)		Performance VALID (%)		Performance TEST (%)	
Center crop	1-99 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	93,86	accuratezza	91,94	accuratezza	87,78
		0	93,70%	5,40%	0	91,20%	5,20%	0	86,40%	9,00%	precision	77,96	precision	72,99	precision	74,39
		1	6,30%	94,60%	1	8,80%	94,80%	1	13,60%	91,00%	specificità	93,68	specificità	91,23	specificità	86,36
											Gmean	94,15	Gmean	92,99	Gmean	88,67
	1-99 percentile	Train			Valid			Test			Performance TRAIN (%)		Performance VALID (%)		Performance TEST (%)	
		Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	95,52	accuratezza	95,67	accuratezza	87,03
		0	96,30%	7,60%	0	96,10%	6,00%	0	88,10%	15,40%	precision	85,39	precision	85,71	precision	75,56
		1	3,70%	92,40%	1	3,90%	94,00%	1	11,90%	84,60%	specificità	96,27	specificità	96,08	specificità	88,1
											Gmean	94,29	Gmean	95,05	Gmean	86,32

Tabella 11 Performance sulle ulteriori due prove di ottimizzazione per la GoogleNet.

Per quanto riguarda la prima prova, come da aspettative, si nota un leggero underfitting su Training e Validation Set che porta alla diminuzione del gap di accuratezza tra Training/Valid e Test Set. L'accuratezza complessiva del Test rimane pressoché costante, con un valore dell'87,78% in questa prova rispetto all'87,63% della prova sulla rete 1-99 crop iniziale. Ciò che migliora notevolmente è la corretta classificazione degli elementi di classe 1 che da una percentuale di 87,60 passa a 91. Quindi a scapito di un leggero peggioramento nella classificazione degli elementi di classe 0 (da 87,7% a 86,4%)

si apprezza il miglioramento nelle performance di classificazione della classe positiva. Per quanto riguarda la seconda prova invece, visti i risultati soddisfacenti portati dalla prima, si è provato ad aumentare ulteriormente il learning rate del layer finale. Questa volta però i risultati non portano a un miglioramento delle performance, infatti, invece di avere un underfitting generale ed equilibrato si ha un underfitting solo per quanto riguarda la classe 1. Inoltre il gap di performance tra Training/Valid e Test set rimane uguale a quello riscontrato nelle prove iniziali.

Tra le due prove quindi si seleziona la prima come migliore e in figura 44 si riporta il grafico per accuratezza e loss function durante la fase di Training di questa rete.

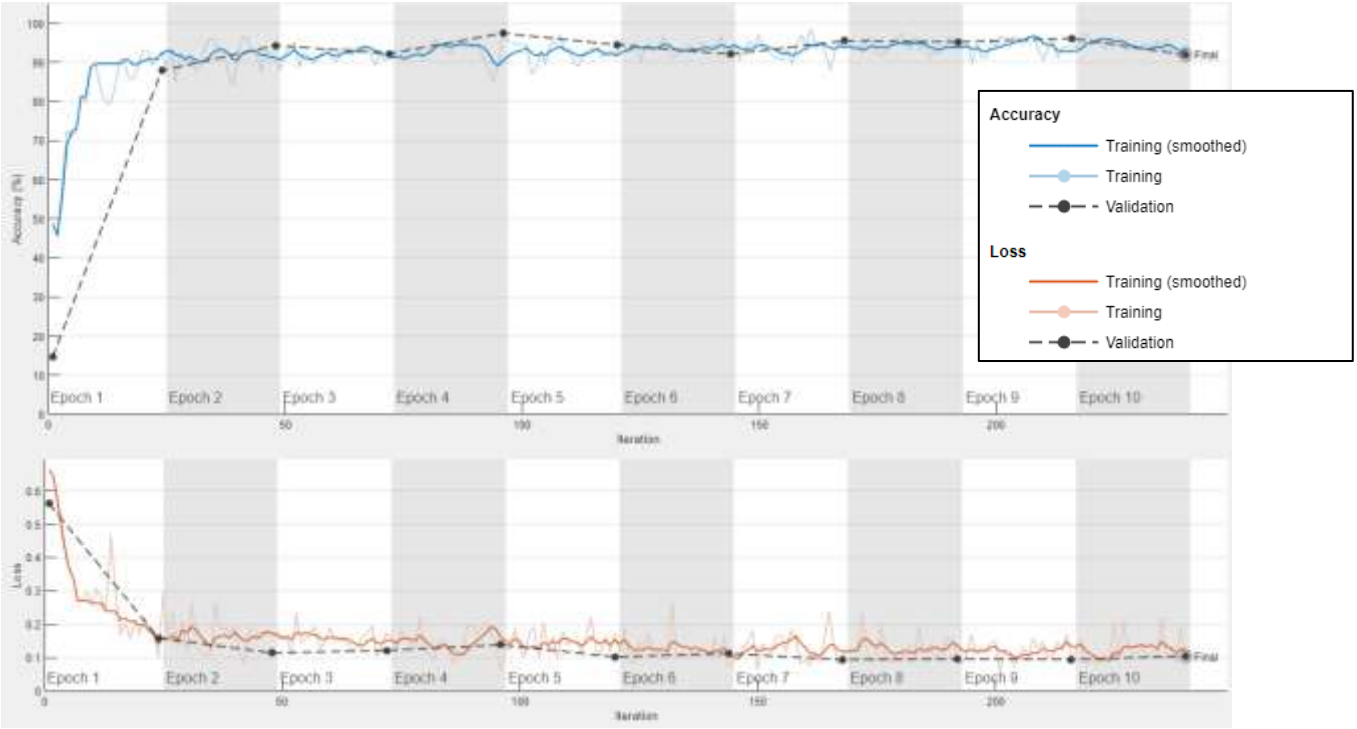

Figura 44 Grafico dell'allenamento della GoogleNet con learning rate del layer finale pari a 50 allenata sul dataset normalizzato con 1-99 percentile e immagini ottenute tecnica center crop.

5.2.2 ResNet

La seconda rete testata è stata la ResNet50. Anche in questo caso sono state eseguite 6 prove combinando 3 tipologie di normalizzazione e due tipologie di crop dell'immagine. I parametri iniziali, uguali per tutte le prove sono stati:

- 150 layers freezzati (è stata utilizzata la tecnica di transfer learning, introdotta al Capitolo 5)
- Learning rate dell' ultimo layer (FullyConnectedLayer) =10
- Learning rate dei layers precedent = 0,001
- 10 epoche di allenamento

		Train			Valid			Test			Performance TRAIN (%)		Performance VALID (%)		Performance TEST (%)	
Center crop	min max	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	95,64	accuratezza	76,12	accuratezza	90,65
		0	96,40%	7,50%	0	73,70%	14,20%	0	96,30%	22,40%	precision	85,85	precision	44,92	precision	90,17
		1	3,60%	92,50%	1	26,30%	85,80%	1	3,70%	77,60%	specificità	96,39	specificità	73,69	specificità	96,32
											Gmean	94,40	Gmean	79,53	Gmean	86,46
	5-95 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	95,96	accuratezza	96,12	accuratezza	89,59
		0	96,40%	6,10%	0	99,10%	15,70%	0	95,70%	24,40%	precision	86,16	precision	95,76	precision	88,37
		1	3,60%	93,90%	1	0,90%	84,30%	1	4,30%	75,60%	specificità	96,44	specificità	99,07	specificità	95,67
											Gmean	95,16	Gmean	91,40	Gmean	85,06
	1-99 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	95,62	accuratezza	96,72	accuratezza	90,80
		0	96,00%	6,00%	0	97,80%	7,50%	0	96,10%	21,40%	precision	84,73	precision	91,18	precision	89,77
		1	4,00%	94,00%	1	2,20%	92,50%	1	3,90%	78,60%	specificità	96,00	specificità	97,76	specificità	96,10
											Gmean	95,00	Gmean	95,11	Gmean	86,92
Center crop + resize	min max	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	96,23	accuratezza	91,34	accuratezza	87,48
		0	96,00%	2,60%	0	90,10%	3,70%	0	87,00%	11,40%	precision	85,03	precision	70,88	precision	74,79
		1	4,00%	97,40%	1	9,90%	96,30%	1	13,00%	88,60%	specificità	95,95	specificità	90,11	specificità	87,01
											Gmean	96,68	Gmean	93,14	Gmean	87,78
	5-95 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	96,86	accuratezza	97,46	accuratezza	87,18
		0	96,70%	2,50%	0	97,40%	2,20%	0	85,70%	9,50%	precision	87,49	precision	90,34	precision	73,39
		1	3,30%	97,50%	1	2,60%	97,80%	1	14,30%	90,50%	specificità	96,71	specificità	97,39	specificità	85,71
											Gmean	97,11	Gmean	97,57	Gmean	88,10
	1-99 percentile	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	96,59	accuratezza	95,97	accuratezza	87,93
		0	96,30%	2,40%	0	96,10%	4,50%	0	85,90%	7,50%	precision	86,30	precision	85,91	precision	74,10
		1	3,70%	97,60%	1	3,90%	95,50%	1	14,10%	92,50%	specificità	96,34	specificità	96,08	specificità	85,93
											Gmean	96,98	Gmean	95,80	Gmean	89,17

Tabella 12 Performance per le prove eseguite sulla rete ResNet.

In tabella 10 sono riassunte tramite confusion matrix e metriche, le performance ottenute con la rete ResNet. Differentemente dalle prove eseguite con la rete GoogleNet, le prove su rete ResNet hanno performato meglio sul Dataset di immagini ottenute con metodica center crop + resize. Infatti, le 3 prove più in basso in tabella 10, riportano si un'accuratezza complessiva più bassa rispetto alle 3 prove più in alto (~87% vs ~ 90%), ma sono preferibili in quanto forniscono una buona classificazione della classe positiva, la più rilevante. Complessivamente, osservando la metrica G-mean, che considera sia sensibilità che specificità, le 3 prove finali sono quelle con migliori performance. Nello specifico l'ultima prova (1-99 percentile e center crop + resize) risulta essere la migliore in assoluto con la più alta percentuale di corretti classificati per la classe 1 sul Test (92,5%) e un buon valore di corretti classificati

per la classe 0 (85,9%). In figura 45 viene riportato l'andamento dell'accuratezza e della funzione loss durante la fase di allenamento della rete. La funzione loss dopo un iniziale netto decremento, oscilla intorno ad un valore di 0,1 stabilizzandosi nelle ultime epoche ad un valore ancora inferiore. L'accuratezza invece, già dalla seconda epoca si porta ad un valore superiore al 95% crescendo leggermente durante la fase di allenamento.

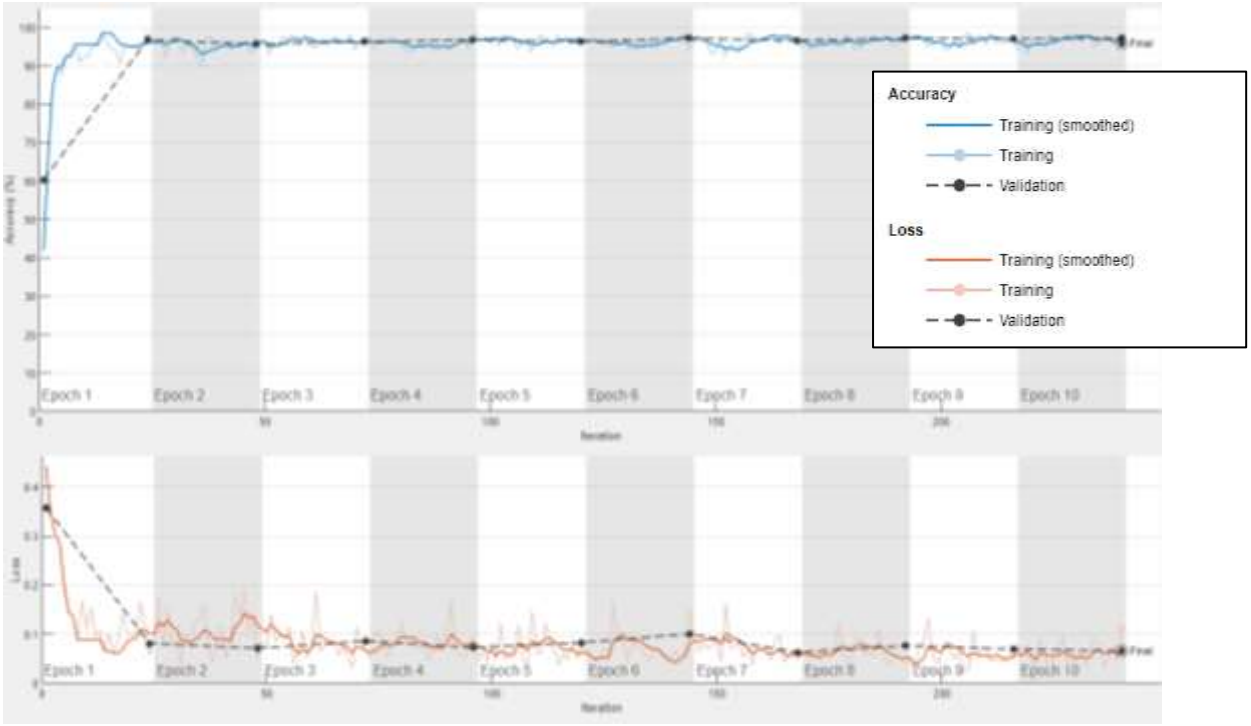

Figura 45 Grafico dell'allenamento della ResNet allenata sul dataset normalizzato con 1-99 percentile e immagini ottenute tecnica center crop + resize.

Individuato allora il miglior Set di immagini per proseguire (normalizzate tra il 1° e il 99° percentile e ottenute con metodica center crop+resize) sono state eseguite ancora alcune prove riducendo i parametri allenabili, rischiando un leggero underfitting, ma con l'obiettivo di diminuire il gap di performance tra Training/Validation Set e Test Set. In particolare, nella prima prova (tabella 11), il learning rate del layer finale è stato aumentato da 10 a 50, nella seconda prova invece il learning rate del layer finale è stato aumentato da 10 a 30 e quello dei layers precedenti è stato modificato da 0,001 a 0,01. Di seguito i risultati ottenuti.

Center crop + resize	1-99 percentile	Train				Valid				Test				Performance TRAIN (%)		Performance VALID (%)		Performance TEST (%)	
		Predetti	Reali			Predetti	Reali			Predetti	Reali			accuratezza	93,69	accuratezza	90,00	accuratezza	82,81
				0	1			0	1			0	1	precision	75,35	precision	67,00	precision	64,36
			0	92,30%	0,50%		0	87,90%	1,50%		0	76,60%	3,00%	specificità	92,32	specificità	87,87	specificità	76,62
			1	7,70%	99,50%		1	12,20%	98,50%		1	23,40%	97,00%	Gmean	95,83	Gmean	93,04	Gmean	86,22
	1-99 percentile	Train				Valid				Test				Performance TRAIN (%)		Performance VALID (%)		Performance TEST (%)	
		Predetti	Reali			Predetti	Reali			Predetti	Reali			accuratezza	98,24	accuratezza	95,07	accuratezza	89,29
				0	1			0	1			0	1	precision	93,82	precision	89,76	precision	78,02
			0	98,50%	2,80%		0	97,60%	14,90%		0	89,00%	10,00%	specificità	98,49	specificità	97,57	specificità	88,96
			1	1,50%	97,20%		1	2,40%	85,10%		1	11,00%	90,00%	Gmean	97,85	Gmean	91,11	Gmean	89,50

Tabella 13 Risultati per le due ulteriori prove di ottimizzazione per la rete ResNet.

Nel primo caso invece di avere un underfitting complessivo si nota un peggioramento delle performance solo per la classe 0. Inoltre, questo peggioramento del modello utilizzato non comporta alcuna riduzione del gap di accuratezza tra la fase di allenamento (93,69%) e di test (82,81%).

Per quanto riguarda la seconda prova i risultati sono più soddisfacenti in quanto si ottengono performance bilanciate tra le due classi e un'accuratezza complessiva sul Test Set pari a 89,29%. Anche questa volta però non migliora il gap di performance tra fase di allenamento e fase di test. Quest'ultima rete ResNet in ogni caso fornisce il miglior compromesso di performance rispetto a tutte le reti ResNet testate. In figura 46 viene riportato l'andamento di accuratezza e loss function durante la fase di allenamento per la rete scelta.

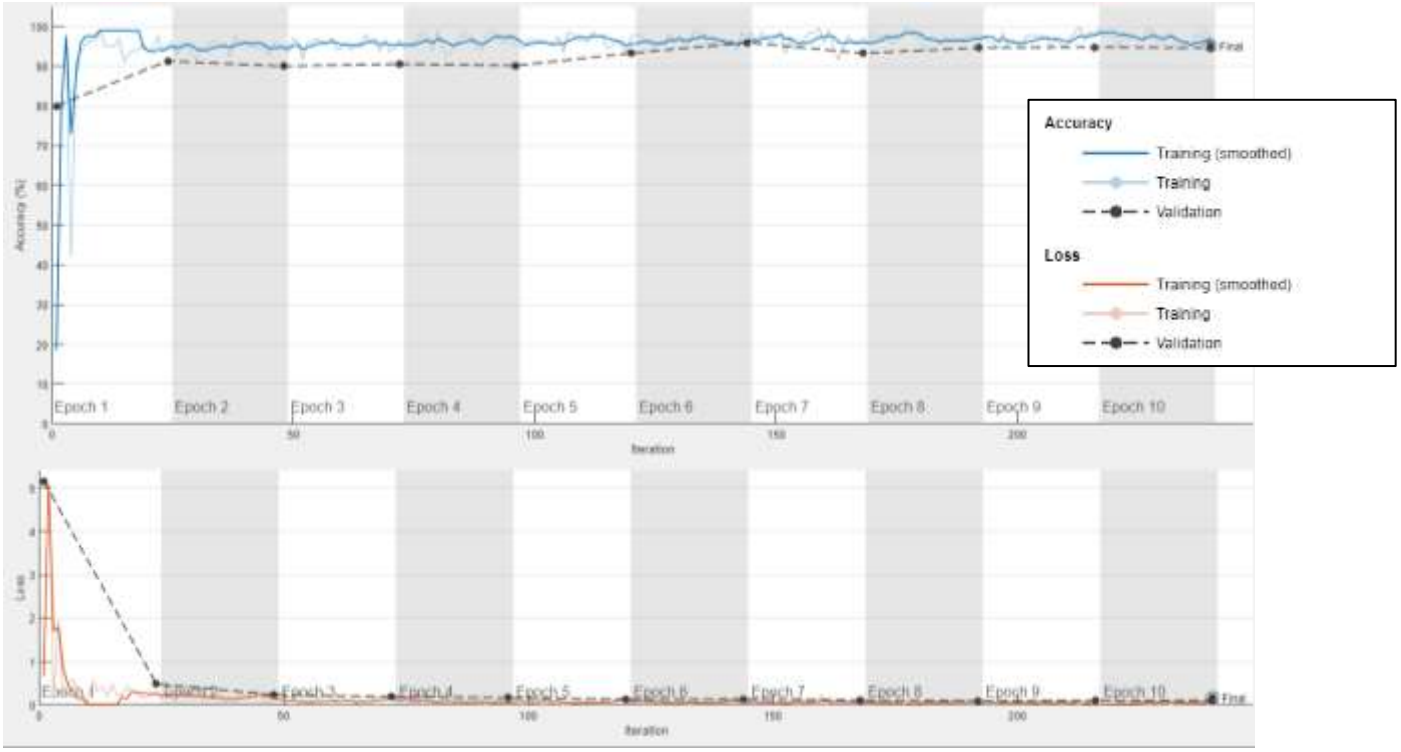

Figura 46 Grafico dell'allenamento della ResNet con learning rate del layer finale pari a 30, learning rate degli altri layer pari a 0,01. La rete è allenata sul dataset normalizzato con 1-99 percentile e immagini ottenute tecnica center crop + resize

5.2.3 Commento finale

Dalle prove effettuate sulle CNN sono emersi risultati soddisfacenti e confrontabili. Si riportano di seguito (tabella 12) le due reti finali scelte, la migliore per ogni tipologia (GoogleNet in azzurro e ResNet in arancione).

GoogleNet	Train			Valid			Test			Performance TRAIN (%		Performance VALID (%		Performance TEST (%)	
	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	93,86	accuratezza	91,94	accuratezza	87,78
	0	93,70%	5,40%	0	91,20%	5,20%	0	86,40%	9,00%	precision	77,96	precision	72,99	precision	74,39
	1	6,30%	94,60%	1	8,80%	94,80%	1	13,60%	91,00%	specificità	93,68	specificità	91,23	specificità	86,36
										Gmean	94,15	Gmean	92,99	Gmean	88,67
ResNet	Train			Valid			Test			Performance TRAIN (%		Performance VALID (%		Performance TEST (%)	
	Predetti \ Reali	0	1	Predetti \ Reali	0	1	Predetti \ Reali	0	1	accuratezza	98,24	accuratezza	95,07	accuratezza	89,29
	0	98,50%	2,80%	0	97,60%	14,90%	0	89,00%	10,00%	precision	93,82	precision	89,76	precision	78,02
	1	1,50%	97,20%	1	2,40%	85,10%	1	11,00%	90,00%	specificità	98,49	specificità	97,57	specificità	88,96
										Gmean	97,85	Gmean	91,11	Gmean	89,50

Tabella 141 Confronto finale tra la miglior rete GoogleNet(in azzurro) e la migliore ResNet (in arancione).

Come già accennato, le performance delle due reti sono confrontabili per quanto riguarda il Test Set. Ciò che risulta migliore nella GoogleNet è il minimo gap tra le performance della fase di allenamento/validazione e la fase di test. In aggiunta nella Google Net si hanno performance confrontabili per le due classi su tutti e 3 i set di dati. Si guardi ad esempio i veri positivi:

- Train: 94,60%
- Valid: 94,80%
- Test: 91%

O i veri negativi:

- Train: 93,70%
- Valid: 91,20%
- Test: 86,4%

Lo stesso non si riscontra nelle confusion matrix delle prove sulla miglior ResNet in cui ad esempio nel Validation Set i veri positivi crollano all'85% rispetto ad un valore di 97% ottenuto sul Train. Per questi motivi si decide di selezionare la GoogleNet come rete ottimale per la risoluzione del problema.

Siccome con le CNN ogni osservazione è riferita ad una singola slice appartenente alla VOI del retto, è necessario aggregare le classificazioni di tutte le fette componenti il volume, ottenute in output dalla GoogleNet, per ricavare la presenza o meno dell'errore di preparazione sulla seduta. Per questo ultimo obiettivo si è scelto di fare un semplice majority voting, ovvero la classe finale del volume è data dalla classe con maggiore co-occorrenza. Nel caso in futuro si potessero avere più pazienti su cui lavorare e soprattutto il feedback di un operatore di radioterapia si potrebbero mettere in atto approcci di post-processing più raffinati.

Conclusioni

Gli enormi passi avanti compiuti dalle tecniche di irradiazione di radioterapia permettono di erogare dosi sempre più elevate ed accurate alla massa tumorale. A fronte di un aumento della probabilità di risposta bisogna però tenere in considerazione il rischio di eventi avversi, monitorati con specifici programmi QA paziente.

Oggigiorno sofisticati sistemi di dosimetria in vivo possono individuare minime discrepanze tra la dose pianificata e quella effettivamente erogata ma non sono in grado di individuare la causa primaria dello scostamento. In questo contesto è nata la necessità di sviluppare un robusto sistema di classificazione dei possibili errori riscontrati durante l'erogazione di radioterapia da parte del software PerFraction di SunNuclear.

Questo studio si proponeva di testare diversi algoritmi per l'analisi di errori limitati ad un solo distretto anatomico. In particolare, l'obiettivo primario era quello di ottimizzare un algoritmo di ML che permettesse di classificare i diversi errori riscontrabili nel distretto della prostata per ogni seduta di radioterapia sulla base di valori quantitativi ottenuti dalle metriche per il confronto tra dose erogata e dose pianificata.

A causa del dataset limitato (20 pazienti per un totale di 152 frazioni) si è deciso di testare classificatori binari per l'individuazione di un solo errore: quello di preparazione del retto per pazienti con carcinoma alla prostata. I classificatori SVM e Multilayer Perceptron hanno fornito risultati leggermente migliori rispetto a quelli di un classificatore random guessing. Queste performance sono giustificabili da diverse problematiche riscontrate durante il corso dello studio: un così ridotto numero di osservazioni non consente di generalizzare il problema, di per sé molto complesso; la distribuzione di dose su cui si allena il classificatore binario è legata anche ad altri possibili errori non tenuti in considerazione o ancor peggio alla sovrapposizione di più errori presenti in quantità differenti. Infine, l'assegnazione del ground truth non eseguita da un operatore esterno di sicuro influenza anch'essa le performance scadenti delle prove eseguite.

In un secondo momento, visti i limiti riscontrati con il primo approccio si è deciso di intraprendere una seconda strada, si è deciso di ripiegare sul metodo classico, usato tuttora per la classificazione di alcuni errori, tra cui quello legato a una scorretta preparazione del retto. Questo metodo consiste nella visualizzazione di immagini CT e CBCT per identificare variazioni anatomiche o errori di setup. In questa fase invece di creare un dataset in cui ogni osservazione è legata ad una frazione, si è costruito un dataset di immagini RGB (composte dalla CT e dalla CBCT) in cui ogni immagine è associata a una slice

di una frazione di un paziente. Queste immagini poi sono state date in pasto alle due reti convoluzionali pre-allenate GoogleNet e ResNet. Questa strada ha restituito risultati molto soddisfacenti.

Per quanto riguarda l'evoluzione del progetto, nel breve termine si svolgerà la validazione della GoogleNet su dati di un secondo centro, per vedere se i risultati ottenuti con questo studio siano generalizzabili. Da questo punto di vista ci si aspettano peggioramenti nelle performance in quanto i valori restituiti dal software PerFraction sono strettamente legati alla calibrazione dei macchinari per radioterapia del centro.

Nonostante il classificatore basato sulle immagini abbia restituito risultati soddisfacenti per l'errore analizzato c'è da tenere in considerazione che alcuni artefatti come quelli legati al movimento degli organi, alla respirazione o alla pulsazione delle arterie non sono identificabili sulla CBCT. Per questo motivo i possibili sviluppi futuri sarebbero principalmente due: in primis estendere lo studio a tutti i distretti anatomici e a tutte le tipologie di errori individuabili dal confronto CT/CBCT. In un secondo momento, nel caso si riescano a collezionare più dati, sarebbe interessante sviluppare un classificatore che integri l'informazione portata da valori dosimetrici con l'informazione portata dalle immagini. In questo modo si potrebbero identificare tutti gli errori presenti in una determinata frazione e magari anche il loro peso sulla distribuzione di dose risultante dopo il trattamento.

Bibliografia e Sitografia

[1] *American Society of Clinical Oncology, "Prostate Cancer." [Online],* Tratto da https://www.cancer.net/cancer-types/prostate-cancer

[2] Hyuna S., Ferlay J., & Siegel R. (4 Febbraio 2021, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries", *ACS Journal*, DOI:10.3322/caac.21660

[3] Fang-zhi C., & Xiao-kun Z., (15 Aprile 2013) "Prostate Cancer: Current Treatment and Prevention Strategies", *PMC*, p. 279–284. DOI:10.5812/ircmj.6499

[4] Botez L., (2017/2018), "*Evaluation of daily dose and identification of sources of error in Volumetric Modulated Arc Therapy treatments with an EPID-based in vivo dosimetry system",* Università degli Studi di Torino- tesi di specializzazione.

[5] National Cancer Institute, "*Neoplasm*", Tratto da: Https://www.cancer.gov/publications/dictionaries/cancer-terms/def/neoplasm

[6] *"La radioterapia oncologica"*, Tratto da https://www.radioterapiaitalia.it/wp-content/uploads/2020/05/6_Scheda_Radioterapia.pdf

[7] National Cancer Institute, "*Curative Radiation Therapy", T*ratto da: https://training.seer.cancer.gov/treatment/radiation/therapy.htm

[8] Webb S.,*The physics of conformal radiotherapy,* Bristol and Philadelphia, Institute of Physics,1997.

[9] Podgorsak E., *Radiation oncology physics: a handbook for teachers and students*, Vienna, International atomic Energy Agency, 2005.

[10] RaiologyInfo*, "Linear Accelerator"*, Tratto da:https://www.radiologyinfo.org/en/info/linac

[11] Alaei P., & Spezi E., (31 Novembre 2015), "Imaging dose from cone beam computed tomography in radiation therapy", *Phys Med*, DOI:10.1016/j.ejmp.2015.06.003

[12] Srinivasan K., Mohammadi M., & Shepherd J., (3 2014 Luglio), "Applications of linac-mounted kilovoltage Cone-beam Computed Tomography in modern radiation therapy", Pubblicato online, DOI:10.12659/PJR.890745

[13] Abi-Jaoudeh N., et al. (25 Luglio 2016), “Prospective Randomized Trial for Image-Guided Biopsy Using Cone-Beam CT Navigation Compared with Conventional CT”, Pubblicato online, DOI:10.1016/j.jvir.2016.05.034

[14] Webb S., *Intensity-modulated radiation therapy.* Bristol and Philadelphia, Institute of Physics, 2001.

[15] Clifford Chao K.S., *Practical Essentials of Intensity Modulated Radiation Therapy,* Philadelphia, Lippincott Williams & Wilkins, 2005.

[16] Enzhuo M., et al.(23 Ottobre 2013), “A comprehensive comparison of IMRT and VMAT plan quality for prostate cancer treatment”, Pubblicato online, DOI: 10.1016/j.ijrobp.2011.09.015.

[17] Bentel G., Nelson C., Noell T. K., *Treatment planning & dose calculation in radiation oncology,* New York, Pergamon Press, 1989.

[18] IAEA, *Accuracy Requirements and Uncertainties in Radiotherapy,* Vienna, International atomic Energy Agency, 2016.

[19] Miri N., Keller P. et al (8 Novembre 2016), “EPID-based dosimetry to verify IMRT planar dose distribution for the aS1200 EPID and FFF beams”, Pubblicato online, *DOI*: 10.1120/jacmp.v17i6.6336

[20] Van Elmpt W., McDermot L., (14 Agosto 2018), “ A literature review of electronic portal imaging for radiotherapy dosimetry”, Pubblicato online, DOI: 10.1016/j.radonc.2008.07.008.

[21] Hussein M., Clark C. (14 Marzo 2017) “Challenges in calculation of the gamma index in radiotherapy - Towards good practice”. Pubblicato online, DOI: 10.1016/j.ejmp.2017.03.001.

[22] “Understanding the Meaning of DVH Metrics”, Tratto da: https://www.carlosjanderson.com/understanding-the-meaning-of-dvh-metrics/

[23] Nalepa J., Kawulok M. (03 Gennaio 2018), “Selecting training sets for support vector machines”, Pubblicato online.

[24] Sahiner B., Pezeshk A., et al. (26 Ottobre 2018), “Deep learning in medical imaging and radiation therapy”, Medical Physics, DOI: https://doi.org/10.1002/mp.13264

[25] “Classification of medical images: understanding the convolutional neural network (CNN)”, Tratto da: https://www.imaios.com/en/Company/blog/Classification-of-medical-images-understanding-the-convolutional-neural-network-CNN

[26] “What Is Transfer Learning? Exploring the Popular Deep Learning Approach2”, Tratto da: https://builtin.com/data-science/transfer-learning

[27] Ho Y., Wookey S., (27 Dicembre 2019) “The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling”, IEEE, DOI: 10.1109/ACCESS.2019.2962617

[28] “A Simple Guide to the Versions of the Inception Network”, Tratto da: https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202

[29] Sudha K.K., Sujatha P., (01 Maggio 2019) “A Qualitative Analysis of Googlenet and Alexnet for Fabric Defect Detection”, International Journal of Recent Technology and Engineering (IJRTE)

[30] Szegedy C., et al., (2015) "Going deeper with convolutions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, DOI: 10.1109/CVPR.2015.7298594.

[31] K. He, X. Zhang, S. Ren and J. Sun, (2016) "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, DOI: 10.1109/CVPR.2016.90.

[32] “Deep Residual Networks (ResNet, ResNet50) – Guide in 2021”, tratto da: https://viso.ai/deep-learning/resnet-residual-neural-network/

[33] “ResNet CNN Networks | Deep Learning Engineer”, tratto da: https://andreaprovino.it/resnet/

[34] “Residual Networks (ResNet) – Deep Learning”, tratto da: https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/

[35] “7 Effective Ways to Deal With a Small Dataset”, tratto da: https://hackernoon.com/7-effective-ways-to-deal-with-a-small-dataset-2gyl407s