Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Matematica

A.a. 2024/2025

Sessione di laurea Novembre/Dicembre 2025

Generazione di Immagini Sintetiche di Melanoma e Nevi Benigni con Generative Adversarial Networks

Relatori:

Dr. Francesco Della Santa

Dott. Paolo Falco

Candidati:

Pietro Rossi

Sommario

Il melanoma è una delle forme tumorali cutanee più gravi e la sua incidenza è in aumento. Poiché il rischio di mortalità cresce in assenza di diagnosi precoce, è fondamentale monitorare nel tempo l'evoluzione dei nei sospetti. In questo contesto, le tecniche di intelligenza artificiale possono offrire un contributo significativo. Questa tesi studia l'impiego delle reti generative avversarie per la creazione di immagini dermoscopiche sintetiche di melanomi e nevi benigni, introducendo elementi innovativi rispetto agli approcci presenti in letteratura.

Dopo un confronto tra diversi modelli generativi, la scelta è ricaduta sulla CycleGAN, che consente di trasformare l'immagine di un nevo benigno nella sua possibile evoluzione melanocitica e, viceversa, di ricostruire il plausibile neo originario dall'immagine di un melanoma reale. La qualità delle immagini generate è stata valutata tramite un sistema di validazione basato su classificatori esterni pre-addestrati (incluso il vincitore della *ISIC 2020 Challenge*) e sulle principali metriche per GAN (*FID*, *KID*, *IS*, *PRDC*), implementate anche con feature estratte da Inception-v3 e ResNet50, riaddestrate su immagini dermoscopiche. Sono stati inoltre sperimentati diversi metodi di sottocampionamento del dataset per valutare come la strategia di selezione dei dati influenzasse la fase di addestramento. A seguito delle analisi delle performance, la rete CycleGAN ha dimostrato la capacità di generare immagini sintetiche di buona qualità per entrambe le classi, simulando in modo convincente l'evoluzione del nevo e la ricostruzione del nevo originario. Tra i diversi metodi, il *cluster sampling rappresentativo* ha prodotto immagini più realistiche e coerenti, pur con una lieve riduzione della variabilità rispetto ad altri metodi.

L'approccio proposto, che integra la costruzione accurata del dataset, un sistema di validazione avanzato e il confronto tra strategie di campionamento, rappresenta un contributo innovativo nel campo della generazione e dell'analisi di immagini di lesioni cutanee, aprendo la strada a futuri sviluppi verso modelli più precisi e clinicamente affidabili.

Ringraziamenti

Vorrei dedicare questa tesi, prima di tutto, alle persone che l'hanno resa possibile e che mi hanno accompagnato passo dopo passo, sia durante il lavoro di implementazione sia nella stesura.

Desidero esprimere la mia profonda gratitudine al professor Della Santa per la sua disponibilità, la vicinanza e l'interesse mostrato verso questo progetto.
Un ringraziamento sincero va anche a tutto l'ambiente aizoOn, per l'esperienza preziosa e per le storie che ciascuna delle persone incontrate mi ha regalato.
In particolare, desidero ringraziare Paolo, che ha creduto in me fin dal primo momento, offrendomi questa opportunità. E, soprattutto, un grazie speciale a Valeria: un'amica, una guida fantastica, attenta, sempre pronta ad ascoltarmi, consigliarmi e accompagnarmi in ogni fase del lavoro e in ogni momento di difficoltà.

Ringrazio poi tutte le persone che ho incontrato durante il mio percorso e che, in un modo o nell'altro, hanno lasciato un segno nella mia vita o mi sono state anche solo per un momento di aiuto, forse inconsapevolmente, nel mio percorso di crescita personale. In particolare:

Tutta la famiglia Fabiani, per la vostra generosità e il vostro cuore immenso, per tutte le carbonare di Piero, per le vacanze indimenticabili al Park Albatros, ma soprattutto per farmi sentire ogni volta parte della vostra famiglia;

Ago, Ube, Rice, Andre, Jeff e Sere, per tutte le belle chiacchierate, mai banali, che mi hanno sempre lasciato uno spunto di riflessione e di crescita personale, e per il legame profondo e speciale che si è creato in questi ultimi anni;

I ragazzi di Cranium, per essere cresciuti insieme a me dal liceo ad oggi, per avermi accompagnato a vivere alcune esperienze che rimarranno indelebili, per esserci stati nei momenti più belli ma anche in quelli più difficili e per essere famiglia ogni volta che ci si rivede nonostante gli anni e le distanze;

Sara, Alessia, Eleonora, Daniele, Matteo, per aver reso speciale letteralmente dal primo giorno il mio percorso universitario, per i viaggi, le esperienze, gli scleri per le sessioni e i progetti di gruppo, per le discussioni sull'organizzazione delle uscite

e per il supporto costante. In particolare Matte, per essere oltre che un amico vero una persona semplice e buona come poche altre, per non avermi mai mandato a quel paese durante la stesura di un paper, per avermi sempre ascoltato e per essere una persona su cui contare in qualsiasi momento;

Lollo, Fers, Fiande, Phil, Tia, Enri, Gianlu, per essere stati le persone più "inaspettate" di questo percorso (in senso positivo ovviamente), per avermi accolto dal primo giorno, da quel ritiro a San Sicario per me ricordo speciale, come se fossi sempre stato parte del vostro gruppo e per avermi permesso di entrare poi a farne parte, per avermi aiutato a crescere pallavolisticamente e caratterialmente, per essere stati a turno i miei taxisti (con menzioni speciali ad Enri e soprattutto a Lollo, con cui i viaggi diventano occasioni per chiacchierare, sfogarsi e scoprire tanta buona musica) e per l'affetto che mi dimostrate e mi date ogni giorno;

Gagu e Virgi, per essere "semplicemente" le persone che dalla prima elementare ad oggi non hanno mai smesso di darmi supporto in ogni cosa, per i pranzi del mercoledì (appuntamento fisso) e i pomeriggi sul divano, per essere una presenza costante nella mia vita ancora oggi, per essere "semplicemente" famiglia;

Riky e Fabi, per essere stati entrambi coinquilini incredibili, per le tantissime chiacchierate, per tutte le serate sbrago, per il supporto reciproco nelle faccende quotidiane e nei momenti di difficoltà, per avermi portato a cambiare tante mie piccole abitudini di vita quotidiana, per aver condiviso e attraversato insieme alcuni dei momenti più significativi della nostra vita. Riky sei stato una conferma, con la tua razionalità e il tuo essere sempre pronto ad ascoltare e ad agire in qualsiasi momento nel bisogno. Fabi invece una bella scoperta, con la tua generosità immensa e la tua capacità di dare consigli di vita sempre sinceri durante il nostro momento pre nanna. Siete persone speciali, di cui sentirò la mancanza;

Infine, tutti i parenti, nonni, zii, cugini, e soprattutto mamma e papà, per essere stati i miei punti di riferimento, per avermi sempre sostenuto in ogni scelta importante, per l'educazione che mi avete dato, per aver messo sempre me davanti ad ogni cosa, nonostante le difficoltà e i diverbi passati: ho sempre pensato che il vostro essere così diversi sia sempre stato un valore aggiunto e non un ostacolo; per esserci ed esserci stati sempre. Siete la cosa più speciale che questa vita mi ha dato.

L'aver ricevuto così tanto affetto mi ha dato tutta la forza di cui avevo bisogno per affrontare questo percorso intenso, e mi piace pensare che l'averlo ricevuto da persone così speciali renda forse un po' speciale anche me.

Indice

Capitolo 1

Introduzione

Il melanoma rappresenta una delle forme più aggressive di tumore cutaneo, caratterizzata da un'elevata capacità di crescita e metastatizzazione se non diagnosticato precocemente. Sebbene costituisca solo una piccola percentuale dei tumori della pelle, è responsabile della maggior parte dei decessi correlati a tali patologie, rendendolo un serio problema di salute pubblica. La diagnosi precoce è quindi cruciale: riconoscere tempestivamente le alterazioni di un nevo può determinare un miglioramento significativo della prognosi del paziente. Tuttavia, distinguere visivamente una lesione benigna da una maligna rimane complesso anche per i dermatologi esperti, a causa dell'ampia variabilità morfologica e cromatica che caratterizza le immagini dermoscopiche [1].

Negli ultimi anni, l'intelligenza artificiale e le reti neurali profonde hanno rivoluzionato la diagnostica dermatologica, dimostrandosi un efficace supporto per l'identificazione automatica di lesioni sospette [1]. Tuttavia, la disponibilità limitata di immagini etichettate e la mancanza di rappresentazioni delle fasi intermedie di evoluzione da nevo a melanoma rappresentano ancora ostacoli significativi all'addestramento di modelli affidabili [2].

In questo contesto, il presente lavoro si propone di generare immagini dermoscopiche sintetiche realistiche di nevi e melanomi mediante l'utilizzo di una rete CycleGAN, con l'obiettivo di esplorare il potenziale di tali immagini sia a fini clinici che formativi. Sebbene le *CycleGAN* siano già state applicate in letteratura a task simili [1, 3], manca ancora un'analisi sistematica e quantitativa che ne valuti in modo approfondito la qualità dei risultati attraverso metriche oggettive e validazione tramite classificatori esterni. Questo studio punta a colmare tale lacuna proponendo un confronto sistematico fra alcune tecniche di raffinamento rispetto allo stato dell'arte per migliorare il dataset di training, rendendo il modello più stabile e accurato e, parallelamente, un approccio di validazione basato su un insieme combinato di metriche quantitative e modelli diagnostici pre-addestrati, in grado

di fornire un riscontro diretto sulla fedeltà e coerenza diagnostica delle immagini generate. Gli studi illustrati sono stato effettuati in *Python*.

Le immagini prodotte al termine di questo lavoro non solo hanno dimostrato un'elevata qualità visiva e morfologica, ma hanno anche raggiunto risultati coerenti con i modelli di riferimento, suggerendo che la rete sia stata in grado di apprendere in modo efficace le caratteristiche discriminanti tra lesioni benigne e maligne.
Questo risultato apre prospettive interessanti: da un lato, può costituire un ulteriore strumento di supporto per i dermatologi, che durante lo screening possono affiancare all'immagine reale del nevo la sua controparte simulata, rendendo più immediata la spiegazione delle differenze tra lesioni benigne e maligne e chiarendo le ragioni per cui un determinato nevo non richiede escissione; dall'altro, può favorire una maggiore consapevolezza nei pazienti, poiché la possibilità di osservare esempi visivi delle potenziali evoluzioni di un nevo rende più immediata la comprensione dei segnali da monitorare. In questo modo, i pazienti sono messi nelle condizioni di riconoscere precocemente eventuali modifiche sospette e di rivolgersi tempestivamente a uno specialista, contribuendo a ridurre il rischio di diagnosi tardive.

La tesi è stata organizzata come segue:

- **Introduzione**: la tesi si apre con una panoramica completa sul tema trattato e una descrizione generale degli obiettivi e del lavoro svolto, così strutturata:
  - **Il melanoma cutaneo**: trattazione che si focalizza sulle caratteristiche cliniche e su alcuni dati rilevanti come l'incidenza, la mortalità e la prevenzione, con un focus sul processo di trasformazione di un nevo benigno in melanoma;
  - **Il progetto IPeR**: descrizione del progetto di ricerca all'interno del quale si colloca questo lavoro;
  - **Intelligenza Artificiale e Applicazioni Dermatologiche in Letteratura**: panoramica della letteratura più recente su tale tematica, con un riepilogo dei metodi e dei risultati più rilevanti che hanno ispirato lo sviluppo del presente studio e con un focus sui modelli generativi avversari (GAN).
  - **Obiettivi e lavoro svolto**: formulazione dettagliata degli obiettivi di ricerca, con una descrizione dello sviluppo del lavoro e con le motivazioni che hanno guidato le scelte metodologiche alla base.
- **Metodologia**: nel secondo capitolo viene descritto in dettaglio l'intero processo sperimentale ed è articolato nelle seguenti sezioni:
  - **Selezione del dataset**: analisi approfondita dei dataset *ISIC 2019* [4] e *ISIC 2020* [5], con particolare attenzione ai metadati, alle procedure

di filtraggio e alle fasi di *preprocessing* adottate per costruire un dataset completo, innovativo e coerente con gli obiettivi del lavoro;

  - **Scelta del modello**: presentazione dei principi teorici alla base del modello *CycleGAN*, con descrizione dettagliata del suo funzionamento e dei vantaggi che la rendono adatta a questa applicazione;
  - **Addestramento del modello**: descrizione dell'architettura implementata, della scelta dei parametri di training e delle principali considerazioni legate alla stabilità e alla convergenza del modello, osservando le metriche interne (loss e discriminatori);
  - **Validazione**: illustrazione delle metriche esterne adottate per la valutazione, insieme alla descrizione dell'utilizzo di classificatori esterni per verificare la coerenza delle immagini sintetiche con quelle reali;
  - **Ottimizzazione dei risultati**: presentazione delle tecniche di sottocampionamento esplorate per la selezione delle immagini di addestramento, del confronto tra le diverse strategie di valutazione e delle scelte finali per l'ottimizzazione dei parametri del modello.

- **Risultati**: analisi dei risultati ottenuti seguendo la metodologia descritta, con valutazioni quantitative e qualitative. La sezione comprende il confronto tra i diversi set di immagini generati mediante le varie strategie di campionamento, la valutazione visiva dei risultati, l'interpretazione dei valori delle metriche e dei classificatori esterni, e la discussione complessiva delle prestazioni del modello finale ottimizzato.

- **Conclusioni**: sintesi dei risultati e prospettive future. La sezione finale riflette sulle criticità emerse, sui possibili sviluppi del modello e sul ruolo delle tecniche generative nella simulazione dei processi oncologici e nelle applicazioni di supporto clinico ed educativo.

Tra i principali *contributi innovativi* di questa tesi si evidenziano la costruzione di un dataset ad hoc corredato da un'analisi dettagliata dei metadati, la definizione di un benchmark completo basato su metriche standard e adattate al task specifico, l'impiego di classificatori esterni per valutare la qualità discriminativa delle immagini sintetiche e, infine, un confronto sistematico tra diverse strategie di raffinamento del training volto a migliorarne stabilità e qualità.

1.1 Il melanoma cutaneo

Il melanoma cutaneo (SKCM, dall'inglese *Skin Cutaneous Melanoma*) è una forma aggressiva di cancro della pelle che origina dai melanociti, le cellule responsabili della produzione di melanina, ovvero il pigmento che conferisce colore alla pelle [6]. Pur rappresentando una quota relativamente ridotta di tutti i tumori cutanei, il melanoma è responsabile di una parte significativa dei decessi correlati ad essi, grazie alla sua elevata capacità di metastatizzare rapidamente ad altri organi [7]. Negli ultimi decenni, l'incidenza del melanoma cutaneo è aumentata, principalmente in correlazione con l'esposizione ai raggi ultravioletti provenienti dalla luce solare e dai dispositivi abbronzanti; citando alcuni dati, si stima che quasi il 90% dei casi sia associato a questi fattori [7]. Inoltre, l'uso di lettini solari, classificati come cancerogeni dall'IARC, aumenta del 75% il rischio di melanoma se utilizzati prima dei 30 anni [8] e subire cinque o più scottature solari tra i 15 e i 20 anni aumenta il rischio di melanoma dell'80% [9].
La diagnosi precoce e un trattamento tempestivo rappresentano elementi cruciali per migliorare le probabilità di successo nella gestione della malattia: si stimano a livello globale circa 57.000 decessi annui a causa del melanoma cutaneo [10]. La mortalità è fortemente influenzata dallo stadio della malattia al momento della diagnosi: risulta più elevata nei pazienti in età avanzata, con metastasi a distanza, con mutazioni genetiche specifiche o in condizioni di immunosoppressione [11]. Nei casi avanzati, la mortalità a cinque anni si aggira tra il 30 e il 40% [11]. Ogni anno si registrano circa 300.000 nuovi casi di melanoma cutaneo a livello mondiale, con un'incidenza particolarmente elevata in Australia, Nuova Zelanda e in alcune aree degli Stati Uniti (California), con circa 25-30 casi ogni 100.000 abitanti; anche in Europa e in Nord America si osserva un incremento costante dei casi [11].
Il melanoma è più frequente nei giovani adulti (20-40 anni), soprattutto nelle donne, mentre negli uomini anziani (60-70 anni) si riscontrano maggiori casi in stadi avanzati [11]. Secondo le proiezioni dell'IARC, tra il 2020 e il 2040 il numero di nuovi casi aumenterà di oltre il 50%, superando i 500.000 all'anno, mentre i decessi correlati al melanoma cresceranno di oltre due terzi [12].

I melanomi cutanei possono essere classificati in quattro forme principali [6]:

- **Melanoma a diffusione superficiale**: è il più comune e meno aggressivo, rappresentando circa il 70% dei casi, e si sviluppa spesso a partire da un neo preesistente.
- **Melanoma nodulare**: è una forma aggressiva, caratterizzata da rapida crescita e alta incidenza di metastasi, e rappresenta circa il 15% dei casi.
- **Melanoma lentigo maligno**: ha un'evoluzione più lenta ed è tipico degli anziani, localizzandosi soprattutto su viso, orecchie e braccia.

- **Melanoma acrale lentiginoso**: si sviluppa su palme, piante dei piedi o sotto le unghie; è più comune in persone con pelle scura e non è correlato all'esposizione solare.

Recenti studi hanno dimostrato che il melanoma può essere classificato anche in sottotipi molecolari, distinti dalle specifiche alterazioni genetiche presenti nelle cellule del melanoma, chiamate mutazioni, la cui identificazione risulta rilevante per la scelta di terapie mirate. La variazione genetica più comune nel melanoma si trova nel gene BRAF [13], che è mutato in circa il 50% dei melanomi cutanei; seguono NRAS (20%), NF-1 (10-15%) e KIT (rispettivamente melanomi acrali, mucosi o lentigo maligno). Il melanoma viene stadiato secondo la classificazione TNM dell'American Joint Committee on Cancer, con stadi da 0 a IV a seconda dell'invasione e della presenza di metastasi [6].

Il rischio di sviluppare melanoma è influenzato da diversi fattori genetici, fenotipici, ambientali e immunologici [6]. Tra i fattori genetici si annoverano la storia familiare di melanoma, la presenza di nevi atipici e mutazioni ereditarie; i fattori fenotipici includono invece carnagione chiara, capelli biondi o rossi, occhi chiari e un'elevata presenza di nei. Tra i fattori ambientali, l'esposizione eccessiva ai raggi ultravioletti e l'uso di lettini solari aumentano significativamente il rischio. Infine, condizioni di immunosoppressione, dovute a terapie o a infezioni come l'HIV, possono favorire lo sviluppo della malattia. L'uso costante di protezioni solari ad alto SPF, l'evitare l'esposizione diretta nelle ore più calde e l'adozione di abbigliamento protettivo costituiscono misure preventive fondamentali. Altrettanto importante è l'auto-monitoraggio della pelle, effettuato mediante l'esame regolare dei nei e l'osservazione di eventuali nuove lesioni [6].
Per stabilire se una lesione meriti escissione, è utile fare riferimento a checklist cliniche o al sistema *ABCDE* [14], che valuta Asimmetria (A), Bordi irregolari (B), Colore irregolare (C), Dimensioni con diametro > 6 mm (D) ed Evoluzione della lesione (E). È importante ricordare che alcuni melanomi, in particolare quelli con lesioni di diametro inferiore a 1 cm, possono non presentare caratteristiche sospette all'ispezione clinica e rivelare la loro natura solo tramite dermatoscopia, che ha notevolmente migliorato la diagnosi precoce delle lesioni melanomatose, consentendo di individuare anche melanomi in fase molto iniziale.

La diagnosi clinica del melanoma dipende fortemente dall'esperienza del medico, anche se il metodo diagnostico certo rimane la *biopsia* della lesione, durante la quale il medico preleva un piccolo campione di tessuto da analizzare in laboratorio [15]. Per i melanomi in stadio avanzato, è spesso necessario completare la valutazione con ulteriori indagini strumentali, come ecografia, tomografia computerizzata (TC), risonanza magnetica (RM) o tomografia a emissione di positroni (PET), talvolta combinata con TC (PET-TC), per identificare eventuali metastasi e determinare le dimensioni del tumore [15]. Il referto istologico, redatto da un patologo o

dermatopatologo, deve fornire informazioni dettagliate sul melanoma, tra cui: lo spessore della lesione, la presenza o assenza di ulcerazione, il tasso mitotico (indicatore della frequenza di divisione cellulare), il tipo o sottotipo di melanoma, la presenza di linfociti infiltranti, lo stato dei margini del campione, la presenza di marcatori associati alla prognosi o alla risposta terapeutica e la presenza di cellule tumorali nei vasi linfatici o sanguigni, nota come invasione linfatica o vascolare. Sulla base di questi dati, possono essere programmati ulteriori test per pazienti ad alto rischio o in stadio avanzato, al fine di definire il miglior percorso terapeutico [15].

Le terapie disponibili per il melanoma includono diverse opzioni, scelte in base allo stadio della malattia e alle caratteristiche molecolari del tumore [16]. La chirurgia rappresenta il trattamento di scelta negli stadi iniziali, con la resezione precoce del melanoma primario. Nei casi più avanzati, l'immunoterapia è spesso impiegata attraverso l'uso di checkpoint inhibitors, anche in combinazione, per migliorare la risposta nei pazienti con malattia avanzata. Le terapie mirate, basate sulle specifiche mutazioni genetiche del tumore, includono inibitori di BRAF e MEK, oltre a inibitori di c-KIT per sottotipi particolari. Infine, la chemioterapia e la radioterapia vengono utilizzate in casi selezionati, quando le altre opzioni non sono efficaci o non applicabili [17].
Negli ultimi anni, strumenti computer-based per la diagnosi hanno consentito una valutazione preliminare delle lesioni sospette utilizzando immagini cliniche di alta qualità, confermate istopatologicamente, e algoritmi di machine learning per la classificazione automatica, permettendo diagnosi più rapide, meno costose e, in molti casi, più accurate [18].

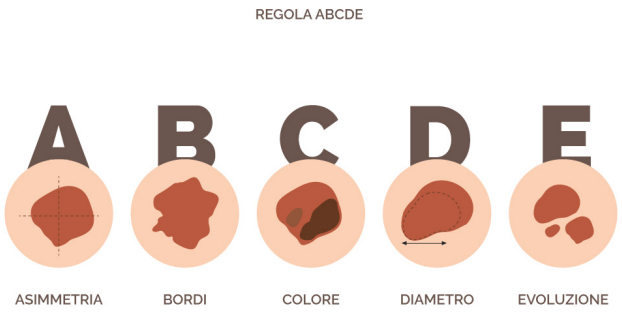


Figura 1.1: Illustrazione della regola diagnostica ABCDE [19].

1.1.1 Trasformazione da nevo benigno a melanoma

La seguente trattazione riprende e riorganizza i principali risultati del lavoro di Shreberk-Hassidim et al. (2022) [20], che analizza in dettaglio le interconnessioni tra nevi e melanomi, con particolare attenzione ai nevi precursori e ai meccanismi di trasformazione.

La trasformazione di un nevo benigno in melanoma è un processo biologico complesso e ancora in parte poco chiaro. Sebbene la maggior parte dei melanomi cutanei insorga *de novo*, circa il 20–30% è associato a un nevo preesistente. I nevi melanocitici, proliferazioni benigne di melanociti, rappresentano infatti uno dei principali elementi di interesse nello studio dei precursori del melanoma. Pur rimanendo nella maggior parte dei casi stabili per tutta la vita, un elevato numero di nevi e la presenza di nevi atipici costituiscono fattori di rischio consolidati: studi epidemiologici hanno infatti dimostrato che le persone con più di cento nevi hanno un rischio circa sette volte maggiore di sviluppare un melanoma, mentre la presenza di numerosi nevi atipici aumenta il rischio di circa sei volte.
Circa un terzo dei melanomi mostra istologicamente una componente nevica residua e viene classificato come *nevus-associated melanoma* (NAM), mentre i restanti casi insorgono *de novo*. La distinzione non è sempre semplice: nei melanomi più spessi, il nevo può risultare mascherato dalla crescita tumorale, oppure possono essere coinvolti micronevi subclinici, troppo piccoli per essere rilevati, ma potenzialmente capaci di agire come lesioni precursori.
La progressione da nevo a melanoma è un fenomeno multistep determinato da una sequenza di eventi molecolari. Mutazioni precoci nei geni *BRAF* o *NRAS* stimolano la proliferazione melanocitaria, portando alla formazione del nevo benigno. In questa fase, le cellule entrano generalmente in uno stato di senescenza che limita l'espansione della lesione: soltanto ulteriori alterazioni genetiche portano alla transizione verso la malignità,come mutazioni del promotore di *TERT*, perdita di geni oncosoppressori e modificazioni di vie di regolazione della crescita e della senescenza. Tra queste, la mutazione *BRAF V600E* è la più frequentemente associata ai nevi acquisiti, mentre le mutazioni *NRAS* si osservano soprattutto nei nevi congeniti di grandi dimensioni. Questo quadro supporta il modello di "evoluzione lineare" dal nevo benigno al melanoma invasivo.
Dal punto di vista clinico e patologico, i NAM presentano caratteristiche distinte rispetto ai melanomi insorti *de novo*. I primi compaiono più frequentemente in età giovanile, in soggetti con un elevato numero di nevi, localizzandosi in aree a esposizione solare intermittente come il tronco; sono spesso melanomi superficiali a diffusione, con spessore ridotto e assenza di ulcerazione, e risultano spesso associati a mutazioni *BRAF V600E*. Al contrario, i melanomi *de novo* tendono a manifestarsi in età più avanzata, su aree cronicamente danneggiate dal sole, come testa e collo, con frequente crescita nodulare, maggiore spessore, presenza di ulcerazione e

alterazioni genetiche differenti, tra cui mutazioni in *TP53*.
Un ulteriore elemento di rilievo è rappresentato dal background pigmentario. Varianti del recettore *MC1R*, tipiche di soggetti con carnagione chiara e capelli rossi, riducono la protezione dai raggi UV e aumentano lo stress ossidativo intracellulare, comportando un rischio più elevato di melanoma indipendentemente dall'esposizione solare. È interessante notare che, pur presentando in media un numero inferiore di nevi rispetto ad altri fototipi chiari, i soggetti con capelli rossi sembrano avere una probabilità maggiore di trasformazione per singolo nevo.

1.2 Il progetto IPeR

L'*Innovative Precision in Oncology Research* (*IPeR*) è un progetto che ha l'obiettivo di sviluppare dispositivi tecnologici e algoritmi per l'oncologia di precisione, con particolare focus sul carcinoma del colon-retto, carcinoma gastrico e melanoma metastatico, patologie che ad oggi sono oggetto di studio su tutti i fronti, sia quello clinico sia tecnologico. Lo scopo finale è semplificare la valutazione e l'implementazione di terapie innovative, migliorando la pratica clinica e personalizzando i trattamenti.
La proposta di IPeR è quella di "creare una catena integrata di analisi cliniche e molecolari che, partendo dalla diagnosi del tumore, porti all'identificazione dei farmaci più promettenti per ulteriori test farmacologici e terapie individualizzate". Questo processo si basa sull'utilizzo degli *organoidi derivati da paziente* (PDO, *Patient Derived Organoids*), modelli biologici in vitro fedeli ai tumori reali, e si sviluppa grazie a tre strumenti tecnologici innovativi:

- **Algoritmi di Intelligenza Artificiale**, impiegati per calcolare il *Polygenic Risk Score*, analizzare i dati clinici e molecolari, e predire la risposta terapeutica;
- **Dispositivi microfluidici di nuova generazione**, progettati per ricreare microambienti tumorali complessi, stimolare la crescita parallela di PDO e monitorarne lo stato in continuo attraverso sensoristica integrata e sistemi di imaging standard;
- **Piattaforme microfluidiche "MicroCATCH"**, dedicate all'isolamento di microvescicole di origine tumorale nel sangue, favorendo diagnosi precoce personalizzate e lo sviluppo di biomarcatori per il monitoraggio della terapia.

Il progetto adotta un approccio interdisciplinare, supportato da dati e campioni raccolti tramite due studi multicentrici retrospettivi e osservazionali, legati al protocollo *PROFILING*, che garantiscono un ampio database temporale per analizzare

endpoint clinici.
L'articolazione del lavoro si sviluppa attraverso quattro fasi principali:

1. Analisi dei dati retrospettivi e *Open Data* con tecniche di Intelligenza Artificiale, per individuare geni associati al rischio di recidiva o resistenza alle terapie;

2. Sviluppo dei dispositivi innovativi per la crescita e il monitoraggio degli organoidi e avvio dei test preliminari della piattaforma *MicroCATCH*;

3. Utilizzo dei PDO per replicare il microambiente tumorale e testare farmaci candidati, con particolare attenzione al concetto di *Biological Twin*;

4. Integrazione dei dati per sviluppare algoritmi predittivi: un *Digital Twin di malattia*, capace di stimare la risposta a diversi trattamenti, e un più ambizioso *Digital Twin di paziente*, mirato a predire la risposta individuale.

IPeR prosegue il percorso avviato dal progetto *DEFLeCT* (focalizzato sul carcinoma polmonare non a piccole cellule, NSCLC) e si integra con il progetto *PNRR D3-4-Health*, estendendone gli obiettivi a nuove patologie e potenziandone le soluzioni tecnologiche. Il coordinamento è affidato a *aizoOn*, società di consulenza tecnologica con consolidata esperienza nella medicina di precisione (si veda Appendice A). All'interno del partenariato, l'IRCCS-FPO fornisce dati e campioni clinici, mentre UniTO e PoliTO apportano competenze biologiche e tecnologiche. Diverse PMI completano il consorzio con know-how specialistico: Fluody per l'analisi chimica, Proplast per dispositivi *lab-on-chip* e BrainDTech per piattaforme microfluidiche dedicate alla *biopsia liquida*.

1.2.1 *OR 1*: Algoritmi per l'oncologia di precisione

La prima componente trasversale dell'innovazione in cui si colloca IPeR è la *Transizione Digitale*, il cui obiettivo principale è l'impiego di algoritmi *Digital Twin*, finalizzati a simulare e prevedere l'evoluzione della patologia, supportando la selezione dei trattamenti terapeutici più adatti. Tali algoritmi operano su diverse scale, dalla simulazione microscopica della malattia alla valutazione del rischio clinico individuale. L'approccio utilizzato è integrato con la disponibilità di dati "Open" e quelli raccolti nell'OR 4, adottando il modello "Data to Action", che privilegia la co-creazione e il miglioramento continuo delle soluzioni attraverso cicli iterativi. In particolare, il progetto affronta tre direttrici principali: l'impiego di algoritmi di *Artificial Intelligence* (AI) e *Machine Learning* (ML), la valorizzazione del dato e della *Data Analysis*, e l'attenzione alla cyber-security.

Sviluppo di algoritmi predittivi: IPeR si propone di progettare modelli di AI e ML per prevedere l'evoluzione della patologia e la risposta ai trattamenti farmacologici. Questi algoritmi saranno integrati con i dispositivi microfluidici sviluppati nel progetto, creando un ecosistema tecnologico che consente di ottenere predizioni cliniche personalizzate. La realizzazione dei *Digital Twin* di malattia e di paziente rappresenta un passo fondamentale verso la medicina personalizzata. Due principi fondamentali guideranno lo sviluppo di questi algoritmi:

- **Intelligenza Artificiale "explainable" e affidabile**: sarà garantita la comprensibilità dei modelli *(explainability)* e la loro affidabilità *(trustworthiness)*, assicurando che i professionisti sanitari possano utilizzare i modelli in modo sicuro e trasparente;
- **Valutazione commerciale**: oltre all'applicazione clinica, gli algoritmi saranno valutati per la loro possibile commercializzazione, con un focus sulla "produttizzazione" per essere utilizzati in contesti di ricerca applicata e, successivamente, in applicazioni cliniche e industriali.

Centralità del dato e medicina data-driven: Uno degli aspetti chiave della Transizione Digitale è la gestione di grandi moli di dati eterogenei provenienti da cartelle cliniche, esami di laboratorio, dati molecolari e organoidi paziente-specifici. L'integrazione di questi dati tramite algoritmi intelligenti costituisce la base per la *Data-Driven Medicine*, che mira a identificare biomarcatori utili per la diagnosi e il trattamento delle patologie oncologiche oggetto di studio.

Cyber-security e protezione dei dati sensibili: La protezione dei dati sanitari è cruciale: pertanto, IPeR integra le migliori pratiche di sicurezza informatica in tutte le fasi del progetto, affrontando specifiche problematiche legate agli attacchi statistici, come *membership inference attacks* e *model inversion attacks*, che mirano a estrarre informazioni sensibili dai modelli di machine learning.

Figura 1.2: Fondazione D34Health

1.3 Intelligenza Artificiale e Applicazioni Dermatologiche in Letteratura

Le GAN sono state originariamente introdotte da Goodfellow et al. (2014) [21], che hanno posto le basi teoriche per la generazione automatica di immagini, mentre la CycleGAN per *unpaired image-to-image translation* è stata proposta da Zhu et al. (2017) [22].
In ambito dermatologico, le applicazioni delle GAN in letteratura possono essere suddivise in quattro categorie principali, come ben evidenziato nella Tabella 1.1: (i) *Data augmentation e classificazione*, la più diffusa [23, 24, 25, 17]; (ii) *Normalizzazione del colore* [26, 27]; (iii) *Valutazione della qualità delle immagini generate* [28, 29]; (iv) *Approcci ibridi e pipeline avanzate* [2].

La diagnosi automatica del melanoma è certamente l'applicazione più studiata e ha ottenuto risultati sempre più accurati grazie allo sviluppo di metodologie avanzate di machine learning. I metodi tradizionali, come Random Forest, Naive Bayes o modelli basati su feature manuali, sono stati progressivamente sostituiti da reti neurali profonde, in particolare Convolutional Neural Networks (CNN), che hanno dimostrato le migliori performance nella classificazione di immagini mediche.
Un contributo rilevante alla comprensione dello stato dell'arte a tal proposito è offerto dalla revisione sistematica di Naseri e Safaei (2025) [18], che analizza 34 studi pubblicati tra il 2016 e il 2024 sull'uso di Machine Learning e Deep Learning per la diagnosi e la prognosi del melanoma da immagini dermoscopiche. Gli autori evidenziano come architetture quali DenseNet, DCNN, ResNet e EfficientNet abbiano raggiunto accuratezze superiori al 95% su dataset benchmark come ISIC e HAM10000, confermando il potenziale delle CNN nell'identificazione precoce del melanoma. La revisione sottolinea inoltre criticità quali la scarsità di dataset bilanciati e annotati, la limitata interpretabilità dei modelli e l'elevato costo computazionale, suggerendo soluzioni come il transfer learning, l'AI spiegabile (es. Grad-CAM, LIME) e la generazione di immagini sintetiche tramite GAN.

Viene proposta una breve analisi di alcuni studi che hanno avuto un impatto sullo sviluppo di questo lavoro di tesi, presentati in ordine temporale, per fornire al lettore una panoramica degli sviluppi progressivi nel campo delle GAN applicate alla dermatologia.

Tra i primi lavori in tale direzione troviamo Rashid et al. (2019) [23] hanno proposto un approccio semi-supervisionato per classificare lesioni cutanee, sfruttando immagini sintetiche GAN per aumentare il dataset ISIC 2018, migliorando le performance rispetto a DenseNet e ResNet-50 fine-tuned e dimostrando il potenziale delle GAN nel superare le limitazioni legate alla scarsità di dati annotati.

Qin et al. (2020) [24] hanno sviluppato SL-StyleGAN, generando immagini dermoscopiche ad alta risoluzione con qualità e diversità superiori ai lavori precedenti, migliorando classificazione e metriche come FID, Precision e Recall.

Salvi et al. (2022) [26] hanno introdotto DermoCC-GAN per la normalizzazione cromatica delle immagini dermatologiche, affrontando il problema della *color constancy* come traduzione immagine-a-immagine, migliorando la qualità dei modelli deep learning e supportando la diagnosi clinica. Successivamente, Salvi et al. (2024) [27] hanno proposto GCC-GAN, basato su dati paired ottimizzati da esperti, migliorando metriche quantitative e robustezza su dataset esterni.

Carrasco Limeros et al. (2022) [28] hanno esplorato l'uso delle GAN per la generazione di immagini di lesioni cutanee, confrontando diversi modelli generativi nel contesto dermatologico. Il loro studio ha evidenziato l'importanza di metriche quantitative come FID e IS per valutare la qualità delle immagini sintetiche e ha proposto approcci per migliorare la diversità e la fedeltà delle immagini generate.

Saeed et al. (2023) [25] hanno combinato GAN ed ESRGAN per generare immagini dermoscopiche ad alta risoluzione e aumentare i dataset ISIC 2019/2020, migliorando la classificazione con CNN e modelli ibridi (VGG19+SVM) fino a un F1-score del 96%.

Wang et al. (2024) [17] propongono un framework in due fasi integrando Condition-StyleGAN2-ADA per generare immagini realistiche di melanoma e MBViT con BatchFormer per la classificazione, raggiungendo un'accuratezza del 98,43% e una sensibilità del 99,01%.

Farady et al. (2024) [2] presentano PSIG-Net, un approccio GAN-based in due fasi volto a gestire ambiguità visiva e outlier: un DC-GAN genera immagini pseudo-realistiche da campioni non ambigui e una rete Siamese valuta la somiglianza, eliminando immagini fuori distribuzione tramite clustering DBSCAN. A differenza di altri approcci GAN-oriented focalizzati sulla generazione massiva, PSIG-Net enfatizza il controllo della qualità dei dati sintetici, raggiungendo accuratezze del 95,8% su ISIC-2017 e 98% su ISIC-2018, superando EfficientNet, ResNet50 e DenseNet121.

Infine, Luschi et al. (2025) [29] propongono un framework completo per la generazione e valutazione di immagini dermoscopiche sintetiche di melanoma maligno, confrontando DCGAN, StyleGAN2 e StyleGAN3-t. DCGAN ha mostrato maggiore diversità, ma con qualità visiva inferiore, mentre StyleGAN2 ottiene i migliori FID (18.89) e KID (0.0025), con valutazioni cliniche cieche da 17 dermatologi che confermano l'indistinguibilità delle immagini sintetiche rispetto alle reali. Il lavoro introduce un protocollo di validazione integrato combinando metriche quantitative (FID, KID, precision, recall) con giudizi esperti, rappresentando un passo fondamentale verso la standardizzazione della valutazione GAN in dermatologia.

Il lavoro da cui questa tesi trae maggiore ispirazione è **Jütte et al. (2024)** [1], che propone un framework innovativo in grado di simulare la progressione delle lesioni cutanee da nevo a melanoma. Il sistema integra una Cycle-GAN con l'interpolazione tra frame e l'analisi del flusso ottico, con l'obiettivo di supportare la diagnosi precoce, la formazione dermatologica e l'educazione del paziente. Il modello è addestrato su 1,571 immagini di nevi e 1,571 di melanomi del dataset SIIM-ISIC, utilizzando una rete generativa basata su ResNet e un discriminatore PatchGAN. L'evoluzione delle lesioni viene valutata attraverso sei metriche derivate dalla regola ABCDE, mostrandone la coerenza: asimmetria (+19%), gradiente del bordo (+63%), convessità (-3%), dispersione cromatica (+45%), diametro (+2%) e variazione cromatica, calcolate mediante tecniche di image processing avanzate; l'analisi del flusso ottico (Farneback) evidenzia l'espansione radiale delle lesioni, correlata all'aumento del diametro, producendo heatmap che identificano le aree di maggiore trasformazione. In aggiunta, un classificatore VGG11 addestrato su HAM10000 mostra una transizione graduale della confidenza diagnostica da nevo a melanoma lungo la sequenza simulata, confermando la coerenza semantica del modello. Il lavoro dimostra come le simulazioni possano essere utilizzate per testare parametri diagnostici, migliorare la comunicazione medico-paziente e supportare protocolli di monitoraggio digitale, rappresentando un contributo metodologico significativo nel campo della dermatologia computazionale.

Rispetto al lavoro di Jütte et al. (2024), questa tesi adotta la stessa configurazione iniziale per la CycleGAN, mantenendo invariati gli iperparametri e utilizzando lo stesso numero di immagini di training per simulare l'evoluzione sintetica in entrambe le direzioni (da nevo a melanoma e viceversa). Tuttavia, l'approccio valutativo si differenzia: mentre gli autori calcolano esplicitamente le metriche ABCDE tramite tecniche di image processing, in questo lavoro tali parametri vengono utilizzati come riferimento visivo preliminare, per poi concentrare la valutazione sulla fedeltà e sul realismo delle immagini generate, attraverso metriche esterne come FID, KID, PRDC e l'impiego di classificatori pre-addestrati. Inoltre, a differenza della selezione casuale delle immagini di training adottata da Jütte, questa tesi esplora e confronta diverse strategie di sottocampionamento, identificando quella più efficace per poi utilizzarla in un training finale esteso, con un numero maggiore di epoche e immagini, al fine di ottimizzare la qualità generativa.

Tabella 1.1: Riepilogo degli studi GAN in dermatologia, suddivisi per tipologia di applicazione.

Categoria	**Autore (anno)**	**Modello GAN**	**Risultati principali**
Data augmentation e classificazione	Rashid et al. (2019)	GAN semi-supervisionato	Migliora DenseNet/ResNet-50 fine-tuned mostrando potenzialità GAN
	Qin et al. (2020)	SL-StyleGAN	Immagini dermoscopiche ad alta risoluzione, maggiore diversità e qualità, miglioramento metriche
	Saeed et al. (2023)	GAN + ESRGAN	Aumento dataset ad alta risoluzione, F1-score 96% con CNN e VGG19+SVM
	Wang et al. (2024)	Condition-StyleGAN2-ADA + MBViT	Immagini generate realistiche, dataset rappresentativo; classificazione con Vision Transformer: accuracy 98.43%, sensitivity 99.01%
Normalizzazione del colore	Salvi et al. (2022, 2024)	DermoCC-GAN, GCC-GAN	Color constancy e standardizzazione cromatica, miglioramento metriche quantitative
Valutazione qualità immagini generate	Carrasco Limeros et al. (2022)	DCGAN, CycleGAN, StyleGAN2, Conditional GAN	Conditional GAN con FID più basso, maggiore somiglianza statistica
	Luschi et al. (2025)	DCGAN, StyleGAN2, StyleGAN3-t	StyleGAN2 con migliori FID (18.89) e KID (0.0025), protocollo di validazione integrato con validazione cieca da dermatologi
Approcci ibridi / pipeline avanzate	Farady et al. (2024)	PSIG-Net (DCGAN + rete Siamese)	Controllo qualità dei dati sintetici, esclusione outlier con DBSCAN, enfasi su qualità generazione; robustezza classificatore, accuracy 95.8–98%

1.3.1 Generative Adversarial Network (GAN)

Una *Generative Adversarial Network (GAN)*, introdotta per la prima volta da Goodfellow et al. (2014) [21], può essere definita come un framework per l'addestramento di modelli generativi basato su un processo avversario. L'idea alla base è quella di mettere in competizione due reti neurali che vengono addestrate simultaneamente: una rete generativa G, il cui scopo è imparare la distribuzione dei dati reali e generare nuovi campioni che la imitino, e una rete discriminativa D, incaricata di distinguere se un campione proviene dal dataset reale oppure è stato prodotto da G.

Durante l'addestramento, il generatore cerca di "ingannare" il discriminatore producendo immagini sempre più realistiche, mentre il discriminatore si aggiorna per migliorare la propria capacità di riconoscere i falsi. Questo meccanismo definisce un gioco minimax a due giocatori, in cui G cerca di massimizzare la probabilità che D commetta un errore e D cerca di minimizzarla. In teoria, esiste un punto di equilibrio unico in cui G riesce a riprodurre perfettamente la distribuzione dei dati reali e il discriminatore non è più in grado di distinguere tra reale e sintetico, assegnando in media una probabilità pari a $\frac{1}{2}$ a entrambi i tipi di campione [21].

Andando più nel dettaglio, e seguendo la trattazione del terzo capitolo di [21], per apprendere la distribuzione del generatore p_g sui dati x si definisce innanzitutto una distribuzione a priori $p_z(z)$ sulle variabili di rumore che fungono da input al modello. Il generatore viene quindi rappresentato come una mappatura dallo spazio del rumore a quello dei dati $G(z;\theta_g)$, dove G è una funzione differenziabile implementata tramite un percettrone multistrato con parametri θ_g.
Parallelamente, si introduce un secondo percettrone multistrato, il discriminatore $D(x;\theta_d)$, che produce in output uno scalare interpretato come la probabilità che un campione x provenga dai dati reali piuttosto che dalla distribuzione generata p_g. Durante l'addestramento, D viene ottimizzato per massimizzare la probabilità di assegnare l'etichetta corretta sia agli esempi reali del dataset, sia ai campioni sintetici generati da G, mentre G viene addestrato per minimizzare la quantità $\log(1 - D(G(z)))$, cercando così di produrre campioni che il discriminatore giudichi sempre più verosimili.

La funzione obiettivo comune è definita come:

$$\min_{\theta_g} \max_{\theta_d} V(D_{\theta_d}, G_{\theta_g}) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D_{\theta_d}(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D_{\theta_d}(G_{\theta_g}(z)))] \tag{1.1}$$

Tale criterio di addestramento consente, in linea teorica, di recuperare la distribuzione generatrice dei dati reali, a condizione che G e D abbiano sufficiente capacità rappresentativa, ossia nel cosiddetto limite non parametrico. Questo significa

che sia il generatore che il discriminatore devono essere in grado di approssimare funzioni altamente complesse, avvicinandosi idealmente a un modello non vincolato dal numero di parametri.

Nella pratica, il gioco deve essere risolto tramite un approccio iterativo e numerico. Un'ottimizzazione completa di D ad ogni passo di training risulterebbe computazionalmente proibitiva e, su dataset finiti, potrebbe portare a fenomeni di *overfitting*. Per tale motivo, l'addestramento viene condotto alternando k passi di aggiornamento del discriminatore a un singolo passo di aggiornamento del generatore, in modo che D rimanga vicino alla propria soluzione ottimale, purché G evolva in maniera sufficientemente graduale. La procedura è formalmente illustrata nell'Algoritmo 1.
Inoltre, nel contesto pratico dell'addestramento, l'Eq. 1.1 potrebbe non fornire un gradiente sufficientemente informativo affinché il generatore G apprenda in modo efficace. Nelle fasi iniziali del training, infatti, quando G è ancora poco performante, il discriminatore D tende a rigettare i campioni sintetici con alta confidenza, poiché essi risultano chiaramente distinti dai dati reali. In tale situazione, il termine $\log(1 - D(G(z)))$ tende a saturare, generando gradienti deboli e rallentando l'apprendimento del generatore.
Per ovviare a questo problema, invece di addestrare G a minimizzare $\log(1 - D(G(z)))$, si può addestrarlo a massimizzare $\log D(G(z))$, così da ottenere lo stesso punto fisso nella dinamica congiunta di G e D, ma con gradienti più intensi nelle fasi iniziali dell'addestramento, migliorando la stabilità e la rapidità di convergenza del modello.

Teoria delle GAN

In linea con quanto discusso nel Capitolo 4 di [21], si approfondiscono alcuni risultati teorici relativi alle GAN ottenuti in un contesto non parametrico, ossia rappresentando un modello con capacità infinita e studiando la convergenza nello spazio delle funzioni di densità di probabilità.
Si riporta che il generatore G definisce implicitamente una distribuzione di probabilità dei dati generati p_g attraverso i campioni $G(z)$, che vengono ottenuti quando z è estratto dalla distribuzione p_z. Pertanto, l'obiettivo è che l'Algoritmo 1 converga verso un buon stimatore di p_{data}, a condizione che il modello abbia sufficiente capacità e tempo di addestramento. Consideriamo prima il discriminatore ottimale D per ogni generatore G.

Proposizione 1. Per un generatore G fissato, il discriminatore ottimale D è:

$$D^*_G(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)} \tag{1.2}$$

Algorithm 1 Addestramento delle reti generative avversarie tramite discesa stocastica del gradiente (minibatch) riportato in [30].

for numero di iterazioni di training **do**
  for k passi **do**
    Campiona un minibatch di m rumori $\{z^{(1)}, \dots, z^{(m)}\}$ dalla distribuzione a priori $p_g(z)$.
    Campiona un minibatch di m esempi $\{x^{(1)}, \dots, x^{(m)}\}$ dalla distribuzione reale dei dati $p_{\text{data}}(x)$.
    Aggiorna il discriminatore ascendendo il suo gradiente stocastico:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D(x^{(i)}) + \log \left(1 - D(G(z^{(i)})) \right) \right].$$

  end for
  Campiona un minibatch di m rumori $\{z^{(1)}, \dots, z^{(m)}\}$ dalla distribuzione a priori $p_g(z)$.
  Aggiorna il generatore discendendo il suo gradiente stocastico:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log \left(1 - D(G(z^{(i)})) \right).$$

end for

Dimostrazione. Il criterio di addestramento per il discriminatore D, dato un generatore G qualsiasi, consiste nel massimizzare la quantità $V(G, D)$:

$$\begin{aligned} V(G, D) &= \int_x p_{\text{data}}(x) \log(D(x)) dx + \int_z p_z(z) \log(1 - D(g(z))) dz \\ &= \int_x p_{\text{data}}(x) \log(D(x)) + p_g(x) \log(1 - D(x)) dx \end{aligned} \tag{1.3}$$

Per ottimizzare questa funzione, notiamo che la funzione $y \to a \log(y) + b \log(1-y)$, per ogni coppia $(a, b) \in \mathbb{R}^2 \setminus \{0,0\}$, raggiunge il massimo in $[0, 1]$ al punto $\frac{a}{a+b}$. Inoltre, il discriminatore non ha bisogno di essere definito al di fuori di $\text{Supp}(p_{\text{data}}) \cup \text{Supp}(p_g)$, per cui, applicando questo risultato alla funzione $V(G, D)$ otteniamo il discriminatore ottimale $D^*_G(x)$. □

Si osservi che il criterio di addestramento per D può essere interpretato come il tentativo di massimizzare il log-likelihood per stimare la probabilità condizionata $P(Y = y \mid x)$, dove Y rappresenta la variabile che indica se il dato x proviene dalla distribuzione dei dati reali p_{data} (con $y = 1$) o dalla distribuzione generata p_g (con

$y = 0$). La formulazione minimax nell'Eq. 1.1 può essere ora riformulata come:

$$C(G) = \max_D V(G, D) = \mathbb{E}_{x\sim p_{\text{data}}} [\log D^*_G(x)] + \mathbb{E}_{z\sim p_z} [\log(1 - D^*_G(G(z)))] \quad (1.4)$$

$$= \mathbb{E}_{x\sim p_{\text{data}}} [\log D^*_G(x)] + \mathbb{E}_{x\sim p_g} [\log(1 - D^*_G(x))]$$

$$= \mathbb{E}_{x\sim p_{\text{data}}} \left[\log \left(\frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}\right)\right] + \mathbb{E}_{x\sim p_g} \left[\log \left(\frac{p_g(x)}{p_{\text{data}}(x) + p_g(x)}\right)\right]$$

Teorema 1. Il minimo globale del criterio di addestramento virtuale $C(G)$ è raggiunto se e solo se $p_g = p_{\text{data}}$. In quel punto, $C(G)$ assume il valore $-\log 4$.

Dimostrazione. Per $p_g = p_{\text{data}}$, il discriminatore ottimale $D^*_G(x) = \frac{1}{2}$ (vedi Eq. 1.2). Perciò, esaminando l'Eq. 1.4 con $D^*_G(x) = \frac{1}{2}$, vediamo che questo è il miglior valore possibile di $C(G)$, raggiunto solo per $p_g = p_{\text{data}}$:

$$C(G) = \log \frac{1}{2} + \log \frac{1}{2} = -\log 4.$$

Osserviamo che:

$$\mathbb{E}_{x\sim p_{\text{data}}}[-\log 2] + \mathbb{E}_{x\sim p_g}[-\log 2] = -\log 4,$$

e sottraendo questa espressione da $C(G) = V(D^*_G, G)$ otteniamo:

$$C(G) = -\log(4) + KL\left(p_{\text{data}} \middle\| \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \middle\| \frac{p_{\text{data}} + p_g}{2}\right),$$

dove KL è la divergenza di Kullback–Leibler. Si riconosce nell'espressione precedente la divergenza di Jensen–Shannon tra la distribuzione del modello e il processo di generazione dei dati:

$$C(G) = -\log(4) + 2 \cdot \text{JSD}(p_{\text{data}} \| p_g).$$

Poiché la divergenza di Jensen–Shannon tra due distribuzioni è sempre non negativa e pari a zero solo se sono uguali, viene così mostrato che $C^* = -\log(4)$ è il minimo globale di $C(G)$ e che l'unica soluzione è $p_g = p_{\text{data}}$, cioè il generatore che replica perfettamente la distribuzione dei dati. □

Proposizione 2. (Convergenza dell'Algoritmo 1) Se G e D hanno capacità sufficienti, ad ogni passo dell'Algoritmo 1 il discriminatore è consentito di raggiungere il suo ottimo dato G e inoltre p_g viene aggiornato per migliorare il criterio:

$$\mathbb{E}_{x\sim p_{\text{data}}}[\log D^*_G(x)] + \mathbb{E}_{x\sim p_g}[\log(1 - D^*_G(x))],$$

allora p_g converge a p_{data}.

Dimostrazione. (Sintesi) Si considera $V(G, D) = U(pg, D)$: poiché $U(p_g, D)$ è convessa in p_g, possiamo applicare il risultato che afferma che il gradiente discendente rispetto a p_g nel punto ottimale di D fornisce gli aggiornamenti corretti. L'Algoritmo 1 garantisce quindi la convergenza di p_g verso p_{data}. □

Nella pratica, tuttavia, le reti avversarie non sono funzioni arbitrarie, ma reti neurali con capacità limitata, che rappresentano solo una classe ristretta di possibili distribuzioni p_g attraverso la funzione $G(z; \theta_g)$. Di conseguenza, invece di ottimizzare direttamente la distribuzione p_g, si ottimizzano i parametri θ_g del generatore. Questo implica che le dimostrazioni teoriche di convergenza non si applicano rigorosamente al caso reale. Nonostante ciò, l'elevata capacità di rappresentazione delle reti neurali profonde osservata empiricamente suggerisce che, pur in assenza di garanzie teoriche formali, il modello sia in grado di approssimare efficacemente la distribuzione dei dati e fornire risultati validi nella pratica.

Infine, nel Capitolo 6 di Goodfellow et al. (2014) vengono esaminati i vantaggi e gli svantaggi delle GAN rispetto ai modelli generativi precedenti. Tra gli svantaggi più rilevanti vi sono l'assenza di una rappresentazione esplicita della distribuzione $p_g(x)$ e la necessità di mantenere una buona sincronizzazione tra il discriminatore D e il generatore G durante l'addestramento, per evitare instabilità.
Sul versante dei vantaggi, le GAN si distinguono per l'efficienza computazionale: non richiedono procedure di campionamento sequenziale tipiche dei metodi basati su catene di Markov, ma si basano esclusivamente sulla backpropagation per l'ottimizzazione dei gradienti. Inoltre, il generatore non apprende copiando direttamente i dati di training, bensì attraverso i segnali forniti dal discriminatore, producendo esempi nuovi e realistici. Un ulteriore beneficio è la capacità di modellare distribuzioni concentrate o anche degeneri, superando i limiti dei metodi tradizionali che necessitano di distribuzioni più "sfocate" per garantire un corretto mescolamento delle modalità [21].

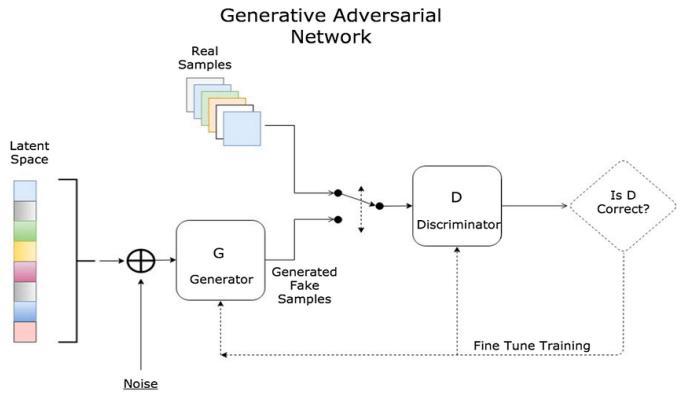


Figura 1.3: L'architettura delle Generative Adversarial Networks [31].

1.4 Obiettivi e Lavoro svolto

L'obiettivo principale di questa tesi è l'utilizzo di modelli generativi basati su reti neurali profonde per la produzione di immagini sintetiche di melanoma realistiche e clinicamente significative. Tra le diverse tipologie di GAN analizzate, è stata scelta la *CycleGAN* per la sua capacità di simulare, a partire da immagini reali, l'evoluzione di un nevo in melanoma e, viceversa, di ricostruire plausibilmente il nevo originario da un melanoma esistente. In particolare, la struttura ciclica del modello garantisce che le immagini generate mantengano la morfologia della lesione di partenza, pur acquisendo le caratteristiche visive tipiche del dominio di destinazione. Inoltre, il fatto che la rete non richieda coppie di immagini corrispondenti rappresenta un vantaggio rilevante, data la scarsità di dataset pubblici contenenti nevi e melanomi "accoppiati".

Dopo la generazione delle lesioni sintetiche e un'attenta analisi dei parametri di valutazione del training per monitorare la stabilità e la convergenza del modello, è stato necessario verificare la qualità delle immagini generate. L'obiettivo non è stato soltanto valutare quanto le immagini di per sé fossero realistiche, ma anche accertare che riproducessero correttamente le caratteristiche dermoscopiche tipiche di nevi e melanomi. Poiché non è stato possibile ricorrere al giudizio diretto di dermatologi esperti, è stato sviluppato un *sistema di verifica* completo, che comprende sia l'ausilio di alcune *metriche esterne consolidate*, comunemente impiegate per valutare le GAN, utili per misurare la qualità e la diversità delle immagini generate, sia l'uso di *classificatori esterni* in grado di distinguere tra lesioni benigne e maligne.
Per rendere la valutazione più aderente al contesto dermoscopico, sono state utilizzate immagini reali di nevi e melanomi per riaddestrare gli ultimi layer delle reti, già pre-addestrate su *ImageNet*, su cui si basano le metriche considerate, al fine di aumentarne la pertinenza rispetto al dominio di interesse.

Successivamente, il lavoro si è concentrato sul *miglioramento della qualità delle immagini* attraverso il raffinamento del dataset di training, con l'implementazione di diversi metodi di sottocampionamento, testando l'influenza di una selezione non casuale delle immagini sulle prestazioni della CycleGAN e permettendo di generare più di una sola possibile evoluzione della stessa lesione. Per il confronto tra le diverse strategie, è stato utilizzato il sistema di validazione sviluppato con lo scopo di individuare le configurazioni più efficaci e aumentare la qualità complessiva delle immagini sintetiche, ottimizzando la scelta dei dati di training. Una volta definita la strategia ottimale, l'obiettivo finale è stato quello di generare immagini della massima qualità possibile, introducendo un'ulteriore ottimizzazione dei parametri di addestramento.

Capitolo 2

Metodologia

In questo capitolo viene descritto il percorso metodologico seguito per la realizzazione del lavoro. Vengono presentate le scelte effettuate in fase di costruzione del dataset, con un analisi dettagliata dei corrispondenti metadati, l'architettura del modello CycleGAN e le modalità di addestramento adottate. Si illustrano inoltre le metriche utilizzate per la validazione, sia nella loro forma standard sia riadattate grazie al fine-tuning delle reti, l'impiego di classificatori esterni per verificare la coerenza delle immagini sintetiche e le strategie di sottocampionamento esplorate per ottimizzare i risultati ed effettuare l'addestramento finale.

2.1 Selezione e preparazione del dataset

In questa sezione viene descritta la fase di selezione e preparazione del dataset utilizzato per l'addestramento del modello generativo. Si introducono innanzitutto i dataset *ISIC* (*International Skin Imaging Collaboration*), illustrandone la natura, gli obiettivi scientifici e il loro ruolo nella ricerca dermatologica.
L'obiettivo è fornire al lettore una visione chiara delle motivazioni che hanno portato alla scelta di queste raccolte di immagini, motivazione che risiede principalmente nella loro ampia accessibilità, nella qualità diagnostica certificata e nella rilevanza clinica per lo studio del melanoma.
Successivamente, l'attenzione viene posta sui dataset scelti, *ISIC 2019* [4] e *ISIC 2020* [5], analizzandone in dettaglio il contenuto e i metadati associati. Viene quindi illustrata la procedura di filtraggio applicata per selezionare solo le immagini pertinenti al task di interesse e, infine, la fase di unione dei due dataset, con l'obiettivo di ottenere un insieme di dati ampio e vario.

2.1.1 I dataset ISIC nella ricerca dermatologica

I dataset dell'*International Skin Imaging Collaboration* (ISIC) [32] rappresentano una risorsa di riferimento per la comunità scientifica nel campo dell'intelligenza artificiale applicata alla diagnostica dermatologica e, in particolare, alla rilevazione precoce del melanoma. Essi contengono decine di migliaia di immagini dermatoscopiche ad alta risoluzione, ciascuna accompagnata da metadati diagnostici e informazioni cliniche di supporto. Le competizioni annuali organizzate dall'ISIC hanno avuto un ruolo determinante nel promuovere l'innovazione nel settore, favorendo lo sviluppo di nuovi algoritmi di classificazione e segmentazione per le lesioni cutanee. Ciò è stato possibile grazie alla disponibilità di dataset standardizzati, costituiti da immagini validate tramite biopsia e corredate da annotazioni di esperti provenienti da centri dermatologici di tutto il mondo. Queste caratteristiche rendono i dataset ISIC strumenti affidabili e di grande valore per la ricerca sulla diagnosi automatica del melanoma e la valutazione dei modelli di apprendimento.

Utilizzo dei dataset ISIC

Come riportato da Hameed et al. (2024) [33], l'utilizzo dei dataset *ISIC* si è rivelato estremamente variegato, con un'attenzione predominante rivolta ai compiti di *classificazione* e *segmentazione* delle lesioni cutanee.
I task di classificazione binaria hanno inizialmente suscitato grande interesse, grazie alla disponibilità di un ampio numero di immagini utili all'addestramento dei modelli di apprendimento automatico, poi consolidato grazie al più ampio e ricco dataset *ISIC 2020*. Con l'introduzione dei dataset *ISIC 2018* e *ISIC 2019*, la ricerca ha progressivamente esteso il proprio raggio d'azione anche verso compiti di classificazione multiclasse.
Al contrario, i compiti di segmentazione non hanno riscosso la stessa popolarità, soprattutto dopo il 2019, anno in cui ISIC ha interrotto le relative challenge. Tuttavia, i dataset pubblicati tra il 2016 e il 2018 mantengono ancora oggi una notevole importanza, poiché includono maschere di segmentazione delle lesioni utili per l'analisi morfologica e la localizzazione precisa delle aree patologiche.
Oltre ai compiti di classificazione e segmentazione, diversi studi hanno esplorato ulteriori direzioni di ricerca basate sui dataset ISIC: tra queste, si annoverano analisi sull'impatto della costanza del colore nelle immagini, nonché l'impiego di tecniche di data augmentation mediante reti generative avversarie, a dimostrazione della versatilità di tali dataset nell'affrontare molteplici aspetti dell'analisi delle immagini dermoscopiche.
Grazie alla loro ricchezza e diversità, i dataset ISIC hanno permesso ai ricercatori di sperimentare e implementare tecniche all'avanguardia basate su algoritmi di Machine Learning e reti neurali profonde. Questa disponibilità di dati di alta qualità ha contribuito a migliorare la precisione e l'efficacia diagnostica dei modelli,

favorendo al contempo la collaborazione internazionale tra istituzioni accademiche e cliniche, anche grazie alle stesse challenge aperte.

ISIC e melanoma

Nel terzo capitolo di [33] viene approfondito il contributo del dataset *ISIC* nella ricerca sul melanoma e il suo impatto clinico. Come già discusso, la diagnosi precoce riveste un ruolo fondamentale nella gestione del melanoma e, in questo contesto, il dataset ISIC ha avuto un ruolo cruciale nel potenziare la capacità diagnostica automatica e nel supportare strategie di trattamento sempre più tempestive ed efficaci.
In particolare, la *challenge ISIC 2020* si è focalizzata sulla rilevazione del melanoma, segnando un punto di svolta nella direzione della classificazione binaria delle lesioni cutanee e favorendo la diffusione di studi orientati alla distinzione tra lesioni benigne e maligne.
Oltre ai progressi nel campo del Machine Learning, analisi recenti condotte sui dataset ISIC hanno indagato anche i marcatori molecolari associati alla progressione del melanoma, evidenziando potenziali bersagli terapeutici per lo sviluppo di strategie di trattamento personalizzate. Questi risultati sottolineano come il dataset ISIC, oltre a costituire una risorsa fondamentale per la diagnostica per immagini, possa fornire informazioni di valore anche per l'ottimizzazione dei percorsi terapeutici dei pazienti affetti da melanoma.

Criticità intrinseche dei dataset ISIC

In questo paragrafo sono messe in evidenza le principali criticità intrinseche dei dataset *ISIC*, spesso trascurate in letteratura, che possono compromettere l'affidabilità dei modelli sviluppati [34]. Vengono quindi discusse le problematiche più rilevanti, in relazione sia alla natura delle immagini sia alle procedure di etichettatura e gestione dei dati, alla luce delle osservazioni riportate in [34] e [33].

Un primo aspetto riguarda la presenza di immagini duplicate o caratterizzate da elevata similarità visiva, che può favorire fenomeni di *overfitting* e introdurre bias, riducendo la capacità del modello di generalizzare correttamente. Un'ulteriore criticità è rappresentata dallo sbilanciamento delle classi, che influisce in modo significativo sulle performance dei modelli e che, in letteratura, viene spesso mitigato mediante tecniche di *data augmentation* o di campionamento mirato. Entrambe queste problematiche sono state riscontrate durante l'analisi dei dataset ISIC condotta in questo lavoro e saranno approfondite nelle sezioni dedicate.

Un elemento altrettanto rilevante è la qualità delle etichette (*label noise*). È stato infatti osservato che diversi dataset di classificazione per il cancro cutaneo includono immagini non confermate da biopsia, con conseguente *ground truth* imperfetta e

rischio di introduzione di errori sistematici. Alcuni studi hanno dimostrato che modelli addestrati su etichette derivate dal consenso di più dermatologi raggiungono prestazioni superiori, soprattutto quando testati su set etichettati a loro volta da specialisti. È stato inoltre evidenziato che le reti neurali convoluzionali non solo apprendono le caratteristiche cliniche comunemente riconosciute dagli esperti, ma tendono anche a interiorizzare le stesse fonti di errore presenti nel processo decisionale umano, con implicazioni significative per l'affidabilità dei modelli e la loro validazione clinica.

Un ulteriore ambito di attenzione riguarda le *questioni etiche e di privacy*, presupposto essenziale per un utilizzo responsabile dell'intelligenza artificiale e dei dati in ambito medico, inclusi quelli del dataset *ISIC*. Negli ultimi anni, la comunità scientifica ha maturato una crescente consapevolezza della necessità di tutelare la riservatezza dei pazienti e di garantire un consenso informato adeguato, favorendo lo sviluppo di protocolli sicuri e trasparenti per la condivisione dei dati.

Infine, la natura dinamica delle patologie cutanee richiede un continuo aggiornamento dei dataset per preservare la validità clinica dei modelli addestrati. La collaborazione *ISIC* risponde attivamente a questa esigenza attraverso l'espansione costante del proprio archivio, includendo nuovi casi e popolazioni più eterogenee, come mostrato nella Tabella 2.1. Sono inoltre in corso iniziative di raccolta mirate e collaborazioni con istituzioni sanitarie internazionali per colmare le lacune ancora presenti, come la scarsità di immagini relative a sottotipi rari o a specifici gruppi demografici [33].

Tabella 2.1: Evoluzione dei dataset ISIC dal 2016 al 2020

Anno	**Train**	**Test**	**Totale**	**Risoluzione**	**Tipologia problema**
ISIC 2016	900	379	1.279	512 × 512	Binario
ISIC 2017	2.000	600	2.600	Variabile	Binario/Multiclasse
ISIC 2018	10.015	1.512	11.527	600 × 450	Multiclasse
ISIC 2019	25.331	8.238	33.569	1024 × 1024	Multiclasse
ISIC 2020	33.126	10.982	44.108	Variabile	Binario

2.1.2 ISIC 2019 e ISIC 2020

In questa sottosezione vengono presentati e analizzati in dettaglio i due dataset ISIC selezionati per lo svolgimento del lavoro: ISIC 2019 e ISIC 2020. La scelta è ricaduta su queste due edizioni non solo per le ragioni già esposte in precedenza, tra cui l'affidabilità diagnostica e la qualità delle immagini, ma anche perché rappresentano i dataset più recenti e maggiormente utilizzati nella letteratura di Machine Learning applicato alla dermatologia.

Il dataset *ISIC 2019* presenta una struttura di etichette *multiclasse* e include un numero significativo di immagini di melanoma, risultando quindi particolarmente utile per la rappresentazione della classe maligna.
Il dataset *ISIC 2020*, invece, contiene un numero complessivamente più ampio di immagini ma una minore quantità di casi di melanoma; è stato tuttavia progettato specificamente per un *task binario* di classificazione *benigno–melanoma*, coerentemente con gli obiettivi del presente lavoro.

Dataset ISIC 2019: [4] comprende 25.331 immagini di training e 8.238 di test, con una risoluzione pari a 1024 × 1024 pixel. I dati etichettati sono disponibili sia per il set di addestramento sia per quello di test e comprendono le seguenti classi: *melanoma*, *nevo melanocitico*, *carcinoma basocellulare*, *cheratosi attinica*, *cheratosi benigna*, *dermatofibroma*, *lesione vascolare* e *carcinoma a cellule squamose*, come visibile in Figura 2.1 e Tabella 2.2. Sono inoltre presenti metadati relativi al paziente, che includono l'età approssimativa, il sesso, il sito anatomico e l'ID della lesione, specificata per 23.247 immagini e mancante per 2.084, con un totale di 11.848 identificativi unici. Il set di test fornisce metadati analoghi, sebbene più limitati (età, sito anatomico e sesso) [34].
È presente anche un sottoinsieme di 2.074 immagini a risoluzione ridotta, contrassegnate dal suffisso *downsampled*, disponibili esclusivamente nel training set. Tuttavia, né sul sito ufficiale né nella documentazione della challenge è riportata una descrizione dettagliata delle modalità di generazione di tali immagini.

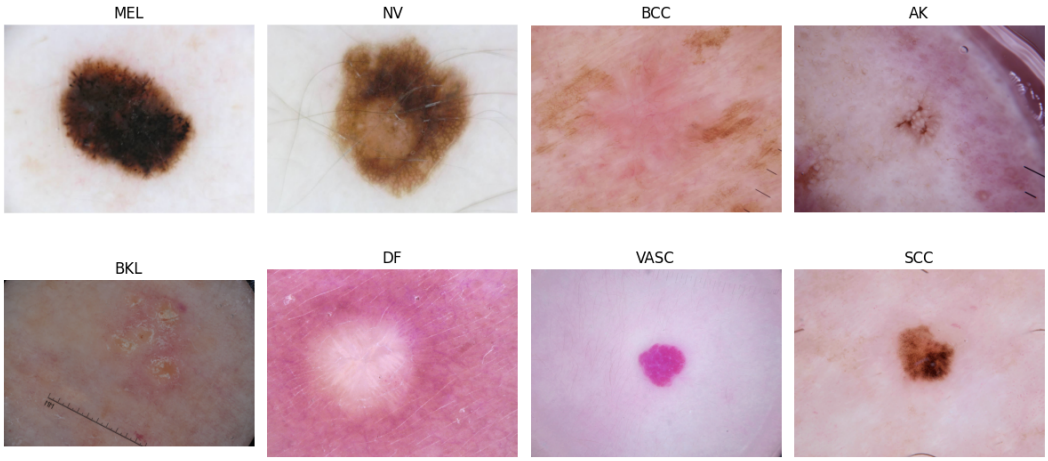


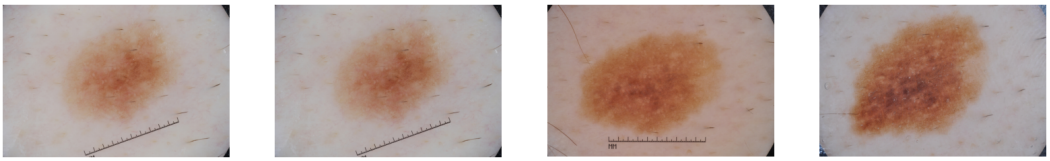
Figura 2.1: Griglia rappresentativa delle otto tipologie di lesioni cutanee presenti nel dataset ISIC 2019: nella prima riga, da sinistra, MEL (*melanoma*), NV (*nevo melanocitico*), BCC (*carcinoma basocellulare*), AC (*cheratosi attinica*); nella seconda riga, BKL (*cheratosi benigna*), DF (*dermatofibroma*), VASC (*lesione vascolare*) e SCC (*carcinoma a cellule squamose*).

Particolare attenzione è stata dedicata all'*ID della lesione*. Tutte le immagini con lo stesso ID (vedi Figura 2.2) presentano sempre la stessa etichetta, il che esclude la possibilità di osservare direttamente una transizione documentata da nevo benigno a melanoma. L'analisi qualitativa ha evidenziato due casi principali: immagini identiche (Fig. 2.2a) e immagini della stessa lesione acquisite da prospettive o condizioni diverse, quali zoom, angolazione, illuminazione (Fig. 2.2b). Quest'ultima caratteristica rappresenta un potenziale valore aggiunto per l'addestramento dei modelli di Deep Learning, poiché consente una rappresentazione più completa delle caratteristiche visive di una lesione e può contribuire a migliorare la robustezza e la capacità di generalizzazione del modello.

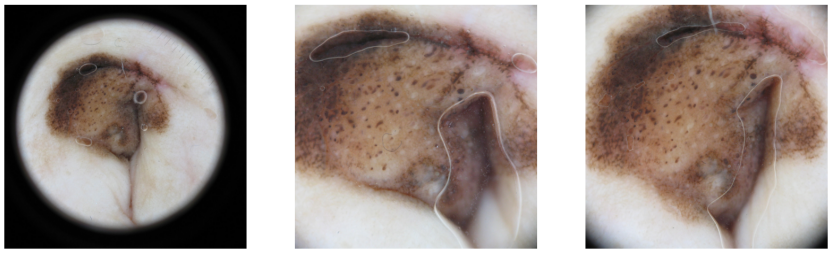
Tabella 2.2: Distribuzione delle classi nel dataset ISIC 2019: MEL (*melanoma*), NV (*nevo melanocitico*), BCC (*carcinoma basocellulare*), AC (*cheratosi attinica*), BKL (*cheratosi benigna*), DF (*dermatofibroma*), VASC (*lesione vascolare*) e SCC (*carcinoma a cellule squamose*), UNK (*unknown).*

Classe	**N° immagini**
MEL	4522
NV	12875
BCC	3323
AK	867
BKL	2624
DF	239
VASC	253
SCC	628
UNK	0

Dataset ISIC 2020: [5] rappresenta, al momento della sua pubblicazione, la collezione più ampia messa a disposizione dalla collaborazione *ISIC*, con 33.126 immagini di training e 10.982 di test, per un totale di 44.108 immagini. Le risoluzioni variano tra i diversi campioni, con una media di circa 768 × 786 pixel. Le etichette sono fornite esclusivamente per il set di addestramento e comprendono informazioni relative all'ID del paziente, all'ID della lesione, al sesso, all'età approssimativa, al sito anatomico, alla diagnosi e allo stato benigno o maligno [34].
Una novità rilevante rispetto alle edizioni precedenti è la presenza dell'*ID paziente* per ogni immagine. Nel training set sono stati identificati 2.056 ID paziente unici e 32.701 ID lesione unici, confermando che numerose immagini provengono da un numero relativamente limitato di pazienti. Questo implica che, in diversi casi, più



(a) ID lesione: `MSK4_0010029`



(b) ID lesione: `BCN_0000010`

Figura 2.2: Esempi di immagini relative alla stessa lesione presenti nel dataset ISIC 2019. Nella figura (a) le prime due immagini a sinistra risultano identiche; tutti gli altri casi mostrano variazioni di acquisizione.

immagini possano riferirsi a lesioni differenti dello stesso paziente, oppure alla stessa lesione acquisita in momenti diversi o con condizioni di illuminazione e angolazione variabili (vedi Figura 2.3).

Un aspetto critico del dataset riguarda la distribuzione delle classi, fortemente sbilanciata: delle 33.126 immagini di training, soltanto 584 sono etichettate come *melanoma*, mentre le restanti 32.542 corrispondono a lesioni benigne. L'analisi della feature *ID lesione* ha inoltre evidenziato che 425 identificativi sono associati a più di un'immagine, di cui soltanto 3 sono riferiti a melanomi. A differenza del dataset *ISIC 2019*, in cui le immagini con lo stesso ID potevano rappresentare viste differenti della stessa lesione, in *ISIC 2020* i duplicati risultano sempre identici e compaiono esclusivamente in coppie (vedi Figura 2.4).
Un'ulteriore variabile di rilievo è la feature *diagnosis*, che specifica la tipologia di lesione e assume i seguenti valori: *unknown*, *nevus*, *melanoma*, *cheratosi seborroica*, *lentigo NOS*, *cheratosi lichenoide*, *lentigo solare*, *macchie caffè-latte* e *proliferazione melanocitaria atipica*. Va tuttavia evidenziata una limitazione significativa: la categoria *unknown* è nettamente predominante, con 27.124 occorrenze, riducendo così l'informatività clinica di questa feature.

Infine, si osserva che per il dataset di test non sono fornite etichette diagnostiche, motivo per cui tale insieme non è stato impiegato nelle analisi. La marcata riduzione del numero di casi di melanoma rispetto a *ISIC 2019*, unita alla presenza di pazienti

ripetuti, costituisce una potenziale fonte di bias, in grado di compromettere la capacità di generalizzazione dei modelli addestrati esclusivamente su questo dataset. D'altra parte, la disponibilità di più immagini relative allo stesso paziente o alla stessa lesione in momenti diversi può offrire indicazioni utili per lo studio della variabilità intra-paziente e dello sviluppo delle lesioni cutanee nel tempo.

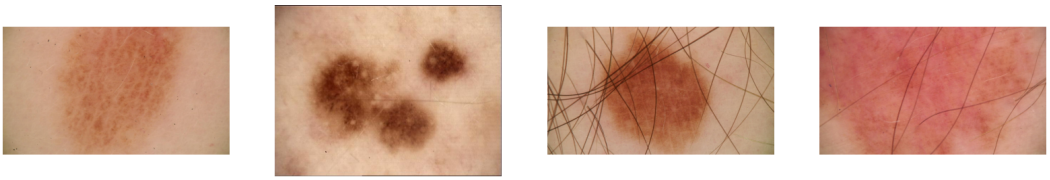
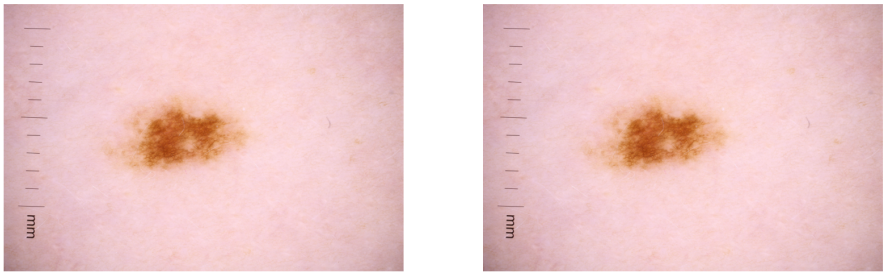
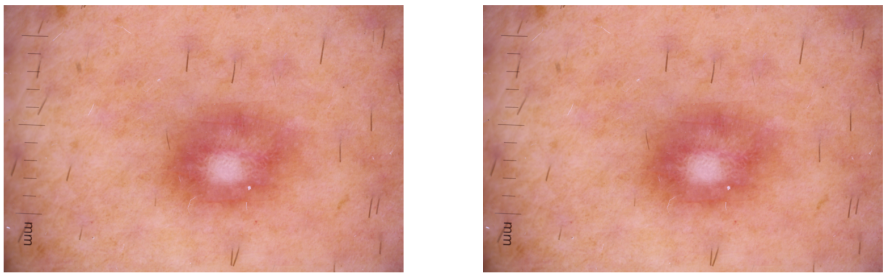


Figura 2.3: Esempi di immagini relative allo stesso ID paziente (`IP_9901629`) presenti nel dataset ISIC 2020.



(a) ID lesione: `IL_9959009`



(b) ID lesione: `IL_9936570`

Figura 2.4: Esempi di immagini relative alla stessa lesione presenti nel dataset ISIC 2020; in entrambe le figure le due coppie risultano identiche.

2.1.3 Preparazione e analisi del dataset finale

A partire da queste considerazioni, si è scelto di utilizzare entrambi i dataset e di combinarli in un unico insieme di dati, al fine di ottenere un dataset finale ampio, bilanciato e coerente con il task di interesse, come definito negli obiettivi 1.4.
La decisione di integrarli nasce dall'intento di combinare i rispettivi punti di forza: da un lato, la maggiore presenza di casi di melanoma in *ISIC 2019*, utile per bilanciare la rappresentazione della classe maligna; dall'altro, l'elevata numerosità di immagini e la struttura binaria di *ISIC 2020*, più coerente con il task *benigno–melanoma*.
L'unione è stata condotta applicando opportune procedure di filtraggio e pulizia, con l'obiettivo di escludere le immagini non pertinenti, come quelle relative ad altri tipi di lesioni cutanee (ad esempio i carcinomi) e garantire così la qualità e l'omogeneità del dataset destinato all'addestramento del modello generativo.
Il risultato di questa procedura è un dataset unificato, costituito da immagini di elevata qualità diagnostica e corredato da metadati completi e coerenti.

2.1.4 Creazione e filtraggio

Poiché l'obiettivo di questo lavoro è simulare l'evoluzione di una lesione in melanoma, è stato necessario individuare, sulla base della letteratura, quali categorie del dataset potessero effettivamente avere una reale possibilità di trasformazione in questa forma tumorale. Dall'analisi di *ISIC 2019* è emerso che, ad eccezione delle classi *nevus melanocitico* e *melanoma*, le altre categorie rappresentano patologie non correlate al melanoma, come carcinomi o lesioni benigne clinicamente distinte. In dettaglio:

- **Carcinoma basocellulare**: tumore maligno distinto dal melanoma;
- **Cheratosi attinica**: forma precoce di carcinoma squamocellulare;
- **Cheratosi benigna**: lesione benigna che non evolve in melanoma, ma talvolta confondibile con esso;
- **Dermatofibroma**: lesione benigna non soggetta a trasformazione maligna, ma in alcuni casi visivamente simile al melanoma;
- **Lesione vascolare**: generalmente benigna e non soggetta a trasformazione melanocitica;
- **Carcinoma a cellule squamose**: tumore maligno distinto dal melanoma;
- **Nessuna delle altre**: categoria residuale, priva di rilevanza clinica per il presente studio.

Alla luce di tali considerazioni, il dataset *ISIC 2019* è stato filtrato mantenendo esclusivamente le classi *nevus melanocitico* e *melanoma* e le etichette sono state quindi rinominate come 0 – benign e 1 – malignant, in modo da uniformarle alla codifica del dataset ISIC 2020. Il set di immagini rimanente comprende 17.397 immagini (4.522 melanomi), mentre il test set contiene 3.822 immagini (1.327 melanomi).

Per coerenza con l'approccio adottato su *ISIC 2019*, sono state escluse in *ISIC 2020* le categorie di lesioni specificate nella variabile *diagnosis* che, secondo la letteratura, non evolvono o non sono correlate al melanoma: *seborrheic keratosis*, *lichenoid keratosis*, *solar lentigo* e *café-au-lait macule*. Dopo tale filtraggio, il dataset risultava composto da 32.946 istanze.

I due dataset così filtrati sono stati infine concatenati al fine di ottenere un insieme di dati più ampio, bilanciato e rappresentativo, che costituisce il punto di partenza per le fasi successive di addestramento e valutazione del modello generativo.

2.1.5 Analisi e considerazioni

L'obiettivo di questa fase è analizzare il dataset unificato per valutarne la completezza, la rappresentatività e l'idoneità rispetto al compito di generazione. Il dataset finale risulta composto da 50.343 istanze, di cui 5.106 etichettate come melanoma e 45.237 come lesioni benigne (vedi Fig. 2.5a).

L'analisi dei valori mancanti mostra che, sul totale delle osservazioni, i *missing values* per ciascuna variabile risultano i seguenti: 470 per l'età approssimata, 2.742 per il sito anatomico, 1.886 per l'ID della lesione e 417 per il sesso. Nel complesso, la percentuale di valori mancanti è contenuta rispetto alla dimensione del dataset e non tale da compromettere l'analisi complessiva. La variabile *ID paziente*, presente unicamente nel dataset *ISIC 2020*, è stata esclusa dai metadati per mantenere una struttura coerente e uniforme.

Sesso: l'analisi della distribuzione delle lesioni per sesso mostra un dataset complessivamente bilanciato, con una leggera prevalenza di pazienti di sesso maschile (25.651 maschi contro 24.275 femmine). Tale equilibrio si mantiene anche considerando separatamente i casi di melanoma e quelli di lesioni benigne, riducendo il rischio di bias legati al genere (vedi Fig. 2.5b.

Area anatomica: l'analisi della distribuzione anatomica delle lesioni evidenzia una maggiore frequenza sul *torso* (16.771) e sugli *arti inferiori* (12.064), seguite dagli *arti superiori* (7.003), dall'*anterior torso* (5.030) e dall'area *testa/collo* (3.457). Meno comuni risultano le lesioni sul *posterior torso* (2.318), su *palmi e piante dei piedi* (744), e rarissime in regioni come *oral/genital* (166) e *lateral torso* (48).

L'analisi incrociata tra area anatomica e tipologia di lesione mostra come in alcune regioni, come il torso, la quasi totalità delle lesioni sia di tipo benigno, mentre in altre (ad esempio *anterior torso* e *head/neck*) la proporzione di melanomi risulta sensibilmente più elevata. Questo suggerisce una diversa incidenza anatomica della patologia, coerente con le osservazioni cliniche in letteratura (vedi Fig. 2.5c).

Età: L'età del paziente rappresenta un fattore discriminante nella probabilità di insorgenza del melanoma. Il range di età coperto dal dataset è compreso tra 0 e 90 anni, con un'età media di 48,9 anni. Dalle statistiche descrittive e dai boxplot emerge che i pazienti con melanoma tendono ad avere un'età superiore rispetto a quelli con lesioni benigne, in accordo con l'evidenza clinica e i dati epidemiologici presenti in letteratura (vedi Fig. 2.5d).

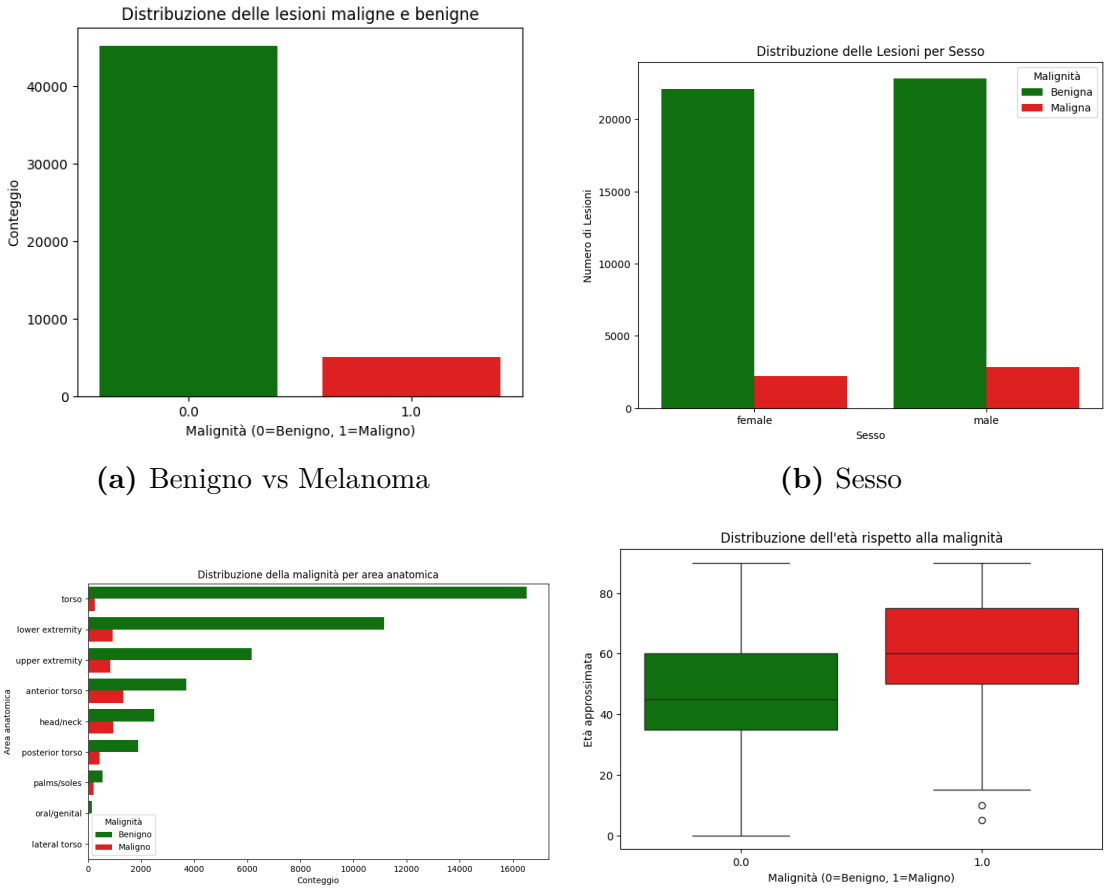
Artefatti e criticità: Durante l'analisi qualitativa delle immagini, sono stati individuati alcuni elementi che possono influire negativamente sulle prestazioni dei modelli di Deep Learning, come già segnalato in [34]. Tra i principali artefatti riscontrati figurano:

- Immagini fortemente ritagliate, con perdita parziale della lesione o dell'area cutanea circostante (vedi Fig. 2.6a);
- Presenza della cornice del campo visivo del dermatoscopio ottico e delle tacche di scala derivanti dallo strumento (vedi Fig. 2.6b);
- Presenza variabile di peli, fattore noto per compromettere l'accuratezza dei modelli (vedi Fig. 2.6c);
- Marcature cliniche attorno alla lesione (vedi Fig. 2.6d);
- Adesivi o righelli fisici di riferimento posti accanto alla lesione (vedi Fig. 2.6e);
- Bolle d'aria dovute all'applicazione del fluido da immersione durante l'esame dermoscopico (vedi Fig. 2.6f).

L'analisi approfondita dei dataset *ISIC* ha permesso di mettere in luce non solo le loro potenzialità, ma anche le criticità intrinseche che possono influenzare in modo significativo l'addestramento dei modelli generativi e che hanno rappresentato un passaggio fondamentale per un impostazione consapevole delle successive fasi del lavoro.

In particolare, la presenza di *immagini duplicate* o caratterizzate da un'elevata similarità visiva può introdurre bias che compromettono la capacità del modello di generalizzare correttamente, determinando previsioni distorte. Questo problema

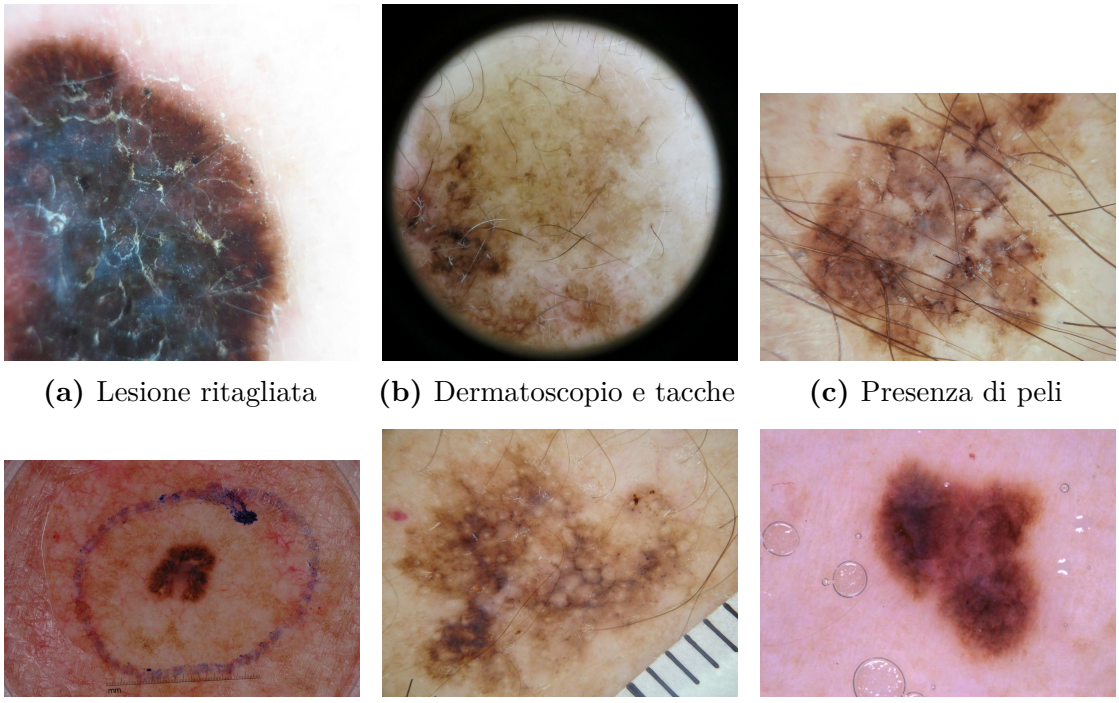
assume un rilievo ancora maggiore quando i dataset *ISIC* vengono ulteriormente ampliati attraverso procedure di *preprocessing* o *data augmentation* [33], come avvenuto nel presente lavoro.
Un'ulteriore criticità è rappresentata dallo *sbilanciamento delle classi*, una condizione comune nei dataset di immagini dermatologiche e che permane anche in seguito all'unione dei due dataset. La letteratura propone diversi approcci per affrontarla, tra cui tecniche di *data augmentation* e di *resampling* volte a garantire una rappresentazione più equilibrata delle diverse tipologie di lesioni.



(a) Benigno vs Melanoma **(b)** Sesso

(c) Area anatomica **(d)** Età

Figura 2.5: Grafici di esplorazione dei metadati del dataset: le barre verdi rappresentano le lesioni benigne, quelle rosse i melanomi. In particolare: (a) distribuzione e bilanciamento delle label benigno–melanoma; seguono, suddivise per tipologia di lesione, (b) età dei pazienti, (c) area anatomica di insorgenza, (d) distribuzione per sesso, .



(a) Lesione ritagliata **(b)** Dermatoscopio e tacche **(c)** Presenza di peli

(d) Tracce di marker **(e)** Scala metrica **(f)** Bolle d'aria

Figura 2.6: Criticità rilevate durante l'analisi qualitativa delle immagini.

2.2 Selezione del modello

Negli ultimi anni, i modelli basati su *Generative Adversarial Networks* (GAN) hanno acquisito una crescente popolarità, specialmente nel campo della sintesi di immagini. Una delle applicazioni più rilevanti delle GAN è l'*image-to-image translation*, ovvero il compito di apprendere una mappatura da un'immagine di input a un'immagine di output. Tradizionalmente, tale processo richiede la disponibilità di un dataset contenente coppie di immagini allineate, in cui a ciascuna immagine di input corrisponde una specifica immagine di output [30, 22].
Sebbene negli ultimi anni siano stati compiuti significativi progressi nella traduzione immagine-a-immagine in contesti supervisionati, persistono alcune limitazioni legate alla disponibilità di dati accoppiati. In molti ambiti, come quello medico, i dataset con immagini appaiate sono estremamente rari; in altri casi, come nello *style transfer* artistico o nella trasformazione di oggetti, tali coppie non sono neppure definibili in modo naturale. Questa scarsità di dati supervisionati riduce la flessibilità e l'applicabilità dei modelli di traduzione tradizionali, motivando lo sviluppo di approcci in grado di operare anche in assenza di corrispondenze dirette tra le immagini dei due domini.

Le *Cycle-Consistent Adversarial Networks* (*CycleGAN*), introdotte da Zhu et al. (2017) [22], offrono una soluzione elegante a tale limitazione. Come descritto anche in [30], a differenza delle GAN standard, nelle quali il generatore viene addestrato per far sì che la distribuzione sintetica segua una distribuzione obiettivo prefissata, la CycleGAN affronta la traduzione tra due *domini non accoppiati* (vedi Figura 2.7). Essa introduce due trasformazioni inverse tra i due domini e impone la *cycle consistency*, ossia la coerenza ciclica, come vincolo fondamentale: l'idea alla base è che, traducendo un'immagine da un dominio all'altro e applicando poi la trasformazione inversa, si dovrebbe poter recuperare l'immagine originale.
In assenza di supervisione diretta, la CycleGAN è in grado di apprendere le caratteristiche distintive di ciascun dominio e di determinare come esse possano essere trasformate per ottenere una corrispondenza con l'altro dominio, rendendo la rete particolarmente adatta a numerosi compiti di traduzione non supervisionata ed efficace anche in contesti con dataset supervisionati limitati.

Per quanto riguarda le applicazioni pratiche delle GAN, Zhu et al. (2017) [22] hanno osservato che la CycleGAN ottiene i migliori risultati nelle trasformazioni che coinvolgono variazioni di *colore e texture*, come la conversione tra diversi stili pittorici, mentre mostra prestazioni inferiori in presenza di *trasformazioni geometriche*.
Analogamente, è stato evidenziato che la *qualità dei risultati* dipende fortemente dalla somiglianza tra le distribuzioni dei due domini: la CycleGAN funziona bene solo quando i domini di partenza e di arrivo condividono una struttura simile. Infine, gli stessi autori di [22] hanno sottolineato che le prestazioni della CycleGAN non possono eguagliare quelle dei modelli basati su dati appaiati e che il modello tende a fallire quando la distribuzione dei dati di training non rappresenta adeguatamente quella dei dati di test. Alcuni esempi di applicazione delle CycleGAN sono mostrati in Figura 2.8.

La scelta di adottare la CycleGAN in questo lavoro, oltre che per le motivazioni teoriche appena evidenziate, deriva dal fatto che essa consente non solo di perseguire l'obiettivo principale della ricerca, ovvero la generazione sintetica di immagini di melanoma realistiche, ma anche di andare oltre e ampliare tale obiettivo in una prospettiva più dinamica: la rete, date le caratteristiche che la contraddistinguono, permette infatti di "tradurre" nel nostro contesto un'immagine di un nevo in una di melanoma e viceversa, offrendo al contempo una *simulazione* dell'evoluzione del nevo verso forme maligne e una *ricostruzione* plausibile del nevo originario a partire da un melanoma reale.

Nelle sezioni successive verranno quindi presentate la formulazione matematica del modello, la trattazione dell'errore e l'analisi dell'excess risk, con l'obiettivo di fornire una panoramica completa e rigorosa dei principi teorici che costituiscono la base della *CycleGAN*.

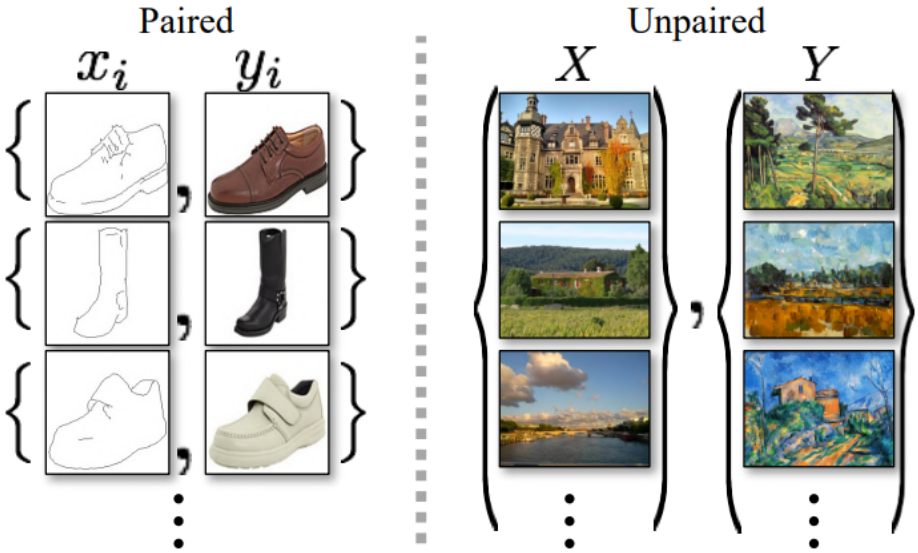


Figura 2.7: Confronto tra dati di training paired (sinistra), in cui ogni esempio x_i ha una corrispondenza diretta con y_i, e dati di training *unpaired* (destra), in cui esistono due insiemi separati (sorgente X e target Y) senza alcuna informazione sulla corrispondenza tra le immagini [22].

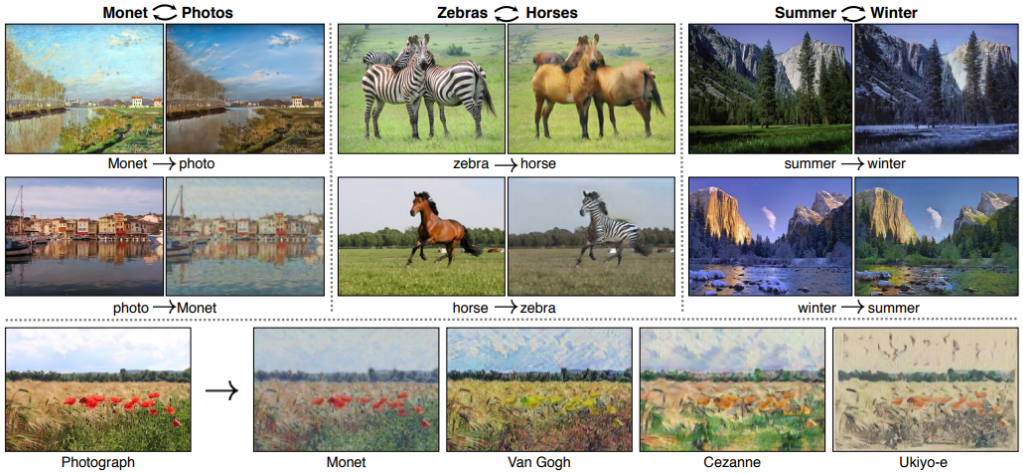


Figura 2.8: Esempi di applicazioni delle CycleGAN: dato un qualsiasi paio di insiemi di immagini non ordinate X e Y, l'algoritmo impara automaticamente a "tradurre" un'immagine da un dominio all'altro e viceversa [22].

2.2.1 Teoria delle CycleGAN

In questa sezione si introduce l'architettura della *CycleGAN*, con l'obiettivo di fornire un quadro teorico chiaro sia del funzionamento del modello sia delle principali proprietà relative alla sua ottimizzazione. Particolare attenzione verrà dedicata all'analisi dell'errore e all'*excess risk*, seguendo l'impostazione proposta in [30].
Nel seguito si assume che i generatori e i discriminatori della CycleGAN siano implementati come reti neurali profonde $\mathcal{NN}(W, \mathcal{L}, B)$ con W larghezza, $\mathcal{L}$ profondità, B vincolo sulla norma e attivazione $ReLU$ (le architetture verranno descritte in dettaglio nelle sezioni successive 2.2.3).

Seguendo la trattazione di [30], la *CycleGAN* applica la proprietà di *cycle consistency* al modello di traduzione combinando due GAN tradizionali in un'unica architettura coerente. Il modello considera due spazi di dati, X e Y, e i corrispondenti spazi delle misure di probabilità $\mathcal{P}(X)$ e $\mathcal{P}(Y)$. La distribuzione di probabilità sui dati reali di X è denotata da $\mu \in \mathcal{P}(X)$, mentre quella su Y da $\nu \in \mathcal{P}(Y)$. L'obiettivo è apprendere due funzioni di traduzione, G e F, tali che le immagini sintetiche prodotte siano distribuite rispettivamente come i domini target ν e μ, considerando le reti neurali generative delle mappe G, F come $\mathcal{NN}(W_G, \mathcal{L}, B_G)$ e $\mathcal{NN}(W_F, \mathcal{L}, B_F)$.
Il processo di generazione diretta è quindi definito come $G : X \to Y$, mentre quello inverso come $F : Y \to X$, con due discriminatori associati, D_X e D_Y; si considerano le reti neurali discriminative delle mappe D_X, D_Y come $\mathcal{NN}(W_{D_X}, \mathcal{L}, 1)$ e $\mathcal{NN}(W_{D_Y}, \mathcal{L}, 1)$ rispettivamente.

La distanza tra la distribuzione generata e quella target viene misurata tramite la *adversarial loss*, calcolata dai discriminatori, mentre la *cycle consistency loss* impone la coerenza tra le mappature diretta e inversa. Siano $F(\nu) \in \mathcal{P}(X)$ e $G(\mu) \in \mathcal{P}(Y)$ le misure *push-forward* delle mappe di traduzione F e G rispettivamente: la distanza tra la distribuzione target μ e la distribuzione generata $F(\nu)$ sotto il discriminatore D_X è definita mediante le *Integral Probability Metrics* (IPM) come:

$$d_{\mathcal{D}_X}(\mu, F(\nu)) = \sup_{D_X \in \mathcal{D}_X} \Big\{ \mathbb{E}_{x\sim\mu}[D_X(x)] - \mathbb{E}_{y\sim\nu}[D_X(F(y))] \Big\},$$

dove $\mathcal{D}_X$ e $\mathcal{D}_Y$ rappresentano le classi di funzioni discriminanti associate ai domini X e Y.

Il problema di ottimizzazione complessivo della CycleGAN è formulato come:

$$\inf_{F\in\mathcal{F},\, G\in\mathcal{G}} L(F, G) = \inf_{F\in\mathcal{F},\, G\in\mathcal{G}} \lambda\, \mathcal{L}_{\text{cyc}}(\mu, \nu, F, G) + d_{\mathcal{D}_X}(\mu, F(\nu)) + d_{\mathcal{D}_Y}(\nu, G(\mu)),$$

dove il parametro $\lambda > 0$ controlla l'importanza relativa della *cycle consistency loss* rispetto all'*adversarial loss*.

In pratica, si utilizzano insiemi finiti di campioni di training $\{x_i\}_{i=1}^n$ estratti da μ e $\{y_j\}_{j=1}^m$ estratti da ν, ottenendo le corrispondenti distribuzioni empiriche $\hat{\mu}$ e $\hat{\nu}$. L'addestramento della CycleGAN risolve quindi il seguente problema di rischio empirico:

$$\inf_{F\in\mathcal{F},\,G\in\mathcal{G}} \hat{L}(F,G) = \inf_{F\in\mathcal{F},\,G\in\mathcal{G}} \lambda\,\mathcal{L}_{\text{cyc}}(\hat{\mu},\hat{\nu},F,G) + d_{\mathcal{D}_X}(\hat{\mu},F_{\#}\hat{\nu}) + d_{\mathcal{D}_Y}(\hat{\nu},G_{\#}\hat{\mu}), \quad (2.1)$$

Poiché l'addestramento opera su distribuzioni empiriche, risulta importante analizzare l'*excess risk* introdotto dal processo di apprendimento. Indicando con $\hat{F},\hat{G}$ la soluzione empirica dell'Eq. (2.1), si definisce l'*excess risk* come:

$$L(\hat{F},\hat{G}) - \inf_{F\in\mathcal{F},\,G\in\mathcal{G}} L(F,G).$$

Si assume che X e Y siano compatti in $\mathbb{R}^d$ con $d \geq 1$, e che le distribuzioni μ e ν siano assolutamente continue. La funzione di perdita complessiva è una combinazione pesata tra *adversarial loss* e *cycle consistency loss*:

$$L(F,G) := \lambda\,\mathcal{L}_{\text{cyc}}(\mu,\nu,F,G) + d_{\mathcal{D}_X}(\mu,F(\nu)) + d_{\mathcal{D}_Y}(\nu,G(\mu)).$$

Le *adversarial losses* per i processi di traduzione diretta e inversa sono definite come:

$$d_{D_X}(\mu,F(\nu)) = \sup_{D_X\in\mathcal{D}_X}\Big\{\mathbb{E}_{x\sim\mu}[D_X(x)] - \mathbb{E}_{y\sim\nu}[D_X(F(y))]\Big\},$$

$$d_{D_Y}(\nu,G(\mu)) = \sup_{D_Y\in\mathcal{D}_Y}\Big\{\mathbb{E}_{y\sim\nu}[D_Y(y)] - \mathbb{E}_{x\sim\mu}[D_Y(G(x))]\Big\}.$$

La *cycle consistency loss* è definita come:

$$L_{\text{cyc}}(\mu,\nu,F,G) := \mathbb{E}_{x\sim\mu}[\|x - F(G(x))\|_1] + \mathbb{E}_{y\sim\nu}[\|y - G(F(y))\|_1],$$

dove $\|\cdot\|_1$ denota la norma ℓ_1.

Supponendo di avere n campioni i.i.d. $\{x_i\}_{i=1}^n$ da μ e m campioni i.i.d. $\{y_j\}_{j=1}^m$ da ν, la funzione di perdita empirica è definita come:

$$\hat{L}(F,G) := \lambda\,L_{\text{cyc}}(\hat{\mu},\hat{\nu},F,G) + d_{D_X}(\hat{\mu},F_{\#}\hat{\nu}) + d_{D_Y}(\hat{\nu},G_{\#}\hat{\mu}),$$

dove $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ e $\hat{\nu} := \frac{1}{m}\sum_{j=1}^m \delta_{y_j}$ rappresentano le distribuzioni empiriche di μ e ν. Nella teoria dell'apprendimento, $L(F,G)$ e $\hat{L}(F,G)$ sono indicati rispettivamente come *expected risk* e *empirical risk*.

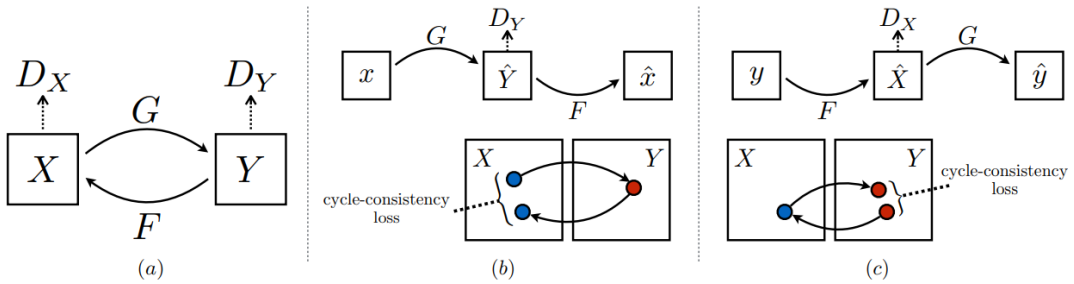


Figura 2.9: Schema del modello CycleGAN: due funzioni di mapping $G : X \to Y$ e $F : Y \to X$, con i rispettivi discriminatori avversari D_Y e D_X. La regolarizzazione è garantita dalle cycle-consistency losses, che impongono coerenza nella traduzione avanti e indietro: (b) forward cycle-consistency $x \to G(x) \to F(G(x)) \approx x$; (c) backward cycle-consistency $y \to F(y) \to G(F(y)) \approx y$ [22].

2.2.2 Analisi dell'errore

La seguente trattazione, basata sul terzo capitolo di [30], introduce il concetto di *excess risk*, che misura la capacità della CycleGAN di generalizzare su dati non osservati. Si considerano i problemi di minimizzazione del rischio atteso ed empirico:

$$\tilde{F}, \tilde{G} := \arg\min_{F,G} L(F, G), \qquad \hat{F}, \hat{G} := \arg\min_{F,G} \hat{L}(F, G),$$

e si definisce l'excess risk come:

$$L(\hat{F}, \hat{G}) - L^*, \qquad L^* := \inf_{F,G} L(F, G),$$

per tutte le funzioni misurabili F, G.

Tale quantità può essere decomposta in due componenti principali:

$$L(\hat{F}, \hat{G}) - L^* = \underbrace{L(\tilde{F}, \tilde{G}) - L^*}_{\text{Errore di approssimazione}} + \underbrace{L(\hat{F}, \hat{G}) - L(\tilde{F}, \tilde{G})}_{\text{Errore di stima}}.$$

La prima componente misura la capacità delle classi di funzioni $\mathcal{F}$ e $\mathcal{G}$ di approssimare la soluzione ottimale, mentre la seconda quantifica l'errore dovuto alla stima empirica basata su un numero finito di campioni.

Per stimare con un upper bound l'*errore di approssimazione*, si sfrutta il legame tra la loss di traduzione e la loss di consistenza ciclica. Il Teorema di Brenier assicura l'esistenza di mappe di trasporto ottimale $\mu \xrightarrow{\nabla\phi} \nu \xrightarrow{\nabla\psi} \mu$, per ϕ, ψ convesse, con $\nabla\phi, \nabla\psi \in H^\alpha$, $\alpha \in (1,2)$. Da questo risultato segue che il rischio ottimale non vincolato soddisfa $L^* = \inf_{F,G} L(F, G) \to 0$, e che rimane da analizzare il termine

$L(\tilde{F}, \tilde{G})$. Il Lemma sull'approssimazione dell'errore di decomposizione mostra che, date funzioni convesse ϕ, ψ,

$$L(F, G) \leq C \sum_{i=1}^{d} \Big(\|\nabla \phi_i - G_i\|_{L^\infty(X)} + \|\nabla \psi_i - F_i\|_{L^\infty(Y)} \Big).$$

Estendendo i risultati di approssimazione a reti ReLU profonde, si riesce ad ottenere che per $\alpha < \frac{d+3}{2}$ e $d > 3$ esistono reti neurali ReLU $f \in \mathcal{NN}(2d+3, \mathcal{L}, B)$ tali che:

$$\sup_{h \in H^\alpha} \|h - f\|_{L^\infty(\Omega)} \lesssim \mathcal{L}^{-\alpha/d} \vee B^{-\frac{2\alpha}{d+3-2\alpha}}.$$

Combinando i risultati discussi, si deriva il tasso di errore di approssimazione per la CycleGAN:

Teorema 1. Siano $X, Y \subseteq [0,1]^d$ con $d > 3$. Esistono reti ReLU F, G con larghezza $W \geq 2d^2 + 3d$, vincolo di norma $B \geq 1$ e profondità $2 \leq \mathcal{L} \leq B^{2d/(d+3-2\alpha)}$ tali che:

$$L(\tilde{F}, \tilde{G}) - L^* \leq O(L^{-\alpha/d}).$$

L'*errore di stima* rappresenta la deviazione tra i generatori empiricamente appresi $(\hat{F}, \hat{G})$ e i generatori ideali $(\tilde{F}, \tilde{G})$. Formalmente, è definito come:

$$L(\hat{F}, \hat{G}) - L(\tilde{F}, \tilde{G}).$$

Esso cattura la discrepanza introdotta dal campionamento finito dei dati e riflette la capacità del modello di generalizzare oltre il dataset di training. Una prima decomposizione mostra che tale errore può essere scritto come:

$$\begin{aligned} L(\hat{F}, \hat{G}) - L(\tilde{F}, \tilde{G}) \leq\ & L(\hat{F}, \hat{G}) - \hat{L}(\hat{F}, \hat{G}) + \hat{L}(\tilde{F}, \tilde{G}) - L(\tilde{F}, \tilde{G}) \\ &= [d_{\mathcal{D}_X}(\mu, \hat{F}_\# \nu) - d_{\mathcal{D}_X}(\hat{\mu}, \hat{F}_\# \hat{\nu}) + d_{\mathcal{D}_X}(\hat{\mu}, \tilde{F}_\# \hat{\nu}) - d_{\mathcal{D}_X}(\mu, \tilde{F}_\# \nu)] \\ &+ [d_{\mathcal{D}_Y}(\nu, \hat{G}_\# \mu) - d_{\mathcal{D}_Y}(\hat{\nu}, \hat{G}_\# \hat{\mu}) + d_{\mathcal{D}_Y}(\hat{\nu}, \tilde{G}_\# \hat{\mu}) - d_{\mathcal{D}_Y}(\nu, \tilde{G}_\# \mu)] \\ &+ \lambda \, [\mathcal{L}_{\text{cyc}}(\mu, \nu, \hat{F}, \hat{G}) - \mathcal{L}_{\text{cyc}}(\hat{\mu}, \hat{\nu}, \hat{F}, \hat{G}) \\ &+ \mathcal{L}_{\text{cyc}}(\hat{\mu}, \hat{\nu}, \tilde{F}, \tilde{G}) - \mathcal{L}_{\text{cyc}}(\mu, \nu, \tilde{F}, \tilde{G})]. \end{aligned}$$

Questa espressione mostra come l'errore di stima sia influenzato da due tipi di errori di stima:

- **Discrepanze di generalizzazione**: differenze tra le distanze misurate con distribuzioni reali (μ, ν) e con le loro versioni empiriche $(\hat{\mu}, \hat{\nu})$: $d_{\mathcal{D}_X}(\mu, \hat{F}_\# \nu) - d_{\mathcal{D}_X}(\hat{\mu}, \hat{F}_\# \hat{\nu})$ e $d_{\mathcal{D}_Y}(\nu, \hat{G}_\# \mu) - d_{\mathcal{D}_Y}(\hat{\nu}, \hat{G}_\# \hat{\mu})$.

- **Consistenza ciclica**: deviazioni tra la loss ciclica calcolata su distribuzioni reali e quella stimata da distribuzioni empiriche: $\mathcal{L}_{\text{cyc}}(\mu, \nu, \hat{F}, \hat{G}) - \mathcal{L}_{\text{cyc}}(\hat{\mu}, \hat{\nu}, \hat{F}, \hat{G})$

Per stimare un limite superiore, si ricorre a strumenti di apprendimento statistico: in particolare la *Rademacher complexity*, che misura la capacità di una classe di funzioni di adattarsi a rumore casuale, e l'integrale di entropia di Dudley. Applicando queste tecniche, si ottiene il seguente risultato:

Teorema 2. Si ritengano valide tutte le assunzioni fatte in precedenza. Poniamo $W := \max\{W_{D_X}, W_{D_Y}, W_F, W_G\}, B := \max\{B_F, B_G\}$. Allora, con probabilità almeno $1 - 12\delta$, vale:

$$L(\hat{F}, \hat{G}) - L(\tilde{F}, \tilde{G}) = O\left(B\Big(\sqrt{\frac{W^2\mathcal{L}}{m}} + \sqrt{\frac{W^2\mathcal{L}}{n}} + \sqrt{\frac{\log(1/\delta)}{m}} + \sqrt{\frac{\log(1/\delta)}{n}}\Big)\right).$$

Questo risultato mostra che l'errore di stima decresce con l'aumentare della numerosità campionaria $N = \max\{m, n\}$ e cresce con la complessità architetturale delle reti (larghezza W, profondità $\mathcal{L}$, vincolo B).

Combinando i risultati precedenti (errore di approssimazione e errore di stima), si ottiene il seguente teorema per un limite superiore dell'excess risk.

Teorema 3. Siano $X, Y \subseteq [0,1]^d$ con $d > 3$ e μ, ν le distribuzioni target su X e Y. Consideriamo n campioni i.i.d. $\{x_i\}_{i=1}^n$ da μ e m campioni i.i.d. $\{y_i\}_{i=1}^m$ da ν. Siano $\mathcal{NN}(W_{D_X}, \mathcal{L}, 1)$ e $\mathcal{NN}(W_{D_Y}, \mathcal{L}, 1)$ le reti dei discriminatori D_X, D_Y, e $\mathcal{NN}(W_F, \mathcal{L}, B_F)$, $\mathcal{NN}(W_G, \mathcal{L}, B_G)$ le reti dei generatori F, G. Definiamo $W := \max\{W_{D_X}, W_{D_Y}, W_F, W_G\}, B := \max\{B_F, B_G\}, N := \max\{m, n\}$. Se le mappe push-forward appartengono alle classi di Hölder H^α con $\alpha \in (1,2)$, allora, con probabilità almeno $1 - 12\delta$, per ogni $N, W \geq 2d^2 + 3d$, quando $B = N^{\frac{d+3-2\alpha}{4d+6}}$, $L = N^{\frac{d}{2d+3}}$, si ha

$$L(\hat{F}, \hat{G}) - L^* \;\leq\; O\Big(N^{-\frac{\alpha}{3+2d}} (\log(1/\delta))^{1/2}\Big).$$

Questo risultato mostra che l'*excess risk* dipende sia dall'errore di approssimazione sia dall'errore di stima. Inoltre, la profondità $\mathcal{L}$ e il vincolo di norma B hanno effetti opposti sui due errori: esiste quindi un punto di equilibrio ottimale, dato da $\mathcal{L} = N^{d/(2d+3)}$ e $B = N^{(d+3-2\alpha)/(4d+6)}$. Infine, la convergenza dell'excess risk presentata suggerisce un quadro di riferimento per la costruzione di reti neurali efficienti nelle CycleGAN, stabilendo una relazione tra la profondità della rete e la dimensione del campione.

2.2.3 L'architettura e le componenti della rete

Come già è stato discusso, l'architettura della CycleGAN si fonda su due componenti fondamentali: i generatori e i discriminatori, che operano in coppia per apprendere una mappatura bidirezionale tra due domini di immagini non corrispondenti.

Il *generatore* è progettato per tradurre un'immagine da un dominio all'altro, preservandone la struttura di base e il contenuto semantico. La sua architettura segue uno schema a tre stadi, tipico delle reti di traduzione di immagine, organizzato come *encoder–transformer–decoder*:

- **Encoder**: nella fase di codifica, uno o più strati convoluzionali con *stride* maggiore di uno eseguono un *downsampling* progressivo dell'immagine. Ciò riduce le dimensioni spaziali e, parallelamente, consente di estrarre rappresentazioni compatte che catturano le principali caratteristiche strutturali e cromatiche dell'input.
- **Transformer**: il nucleo centrale del generatore è costituito da una serie di *residual blocks*, che permettono di apprendere trasformazioni complesse mantenendo la stabilità del gradiente durante il training e, di conseguenza, un flusso stabile delle informazioni. Ogni blocco include un'operazione di convoluzione, seguita da *Instance Normalization* e da un'attivazione *ReLU*.
- **Decoder**: nella fase di decodifica, le feature vengono riportate alla risoluzione originale tramite strati di *upsampling* implementati con convoluzioni trasposte. Anche in questa fase si utilizzano *Instance Normalization* e funzioni di attivazione *ReLU*, mentre l'ultimo layer convoluzionale impiega un'attivazione *tanh* per generare l'immagine tradotta nel dominio target. Per ridurre gli artefatti ai bordi, viene inoltre applicato *reflective padding*.

Questa architettura consente al generatore di produrre immagini realistiche che preservano la struttura globale dell'input, modificandone lo stile e l'aspetto visivo per adattarsi al dominio di destinazione.

Il *discriminatore* della CycleGAN segue invece il paradigma del *PatchGAN discriminator*, introdotto per valutare il realismo delle immagini a livello locale piuttosto che globale. A differenza di un discriminatore tradizionale, che produce un'unica probabilità binaria di "reale" o "falso", il PatchGAN genera una mappa di probabilità in cui ciascun elemento corrisponde a una *patch* dell'immagine di input. Questo approccio consente al modello di concentrarsi sulla coerenza dei dettagli strutturali, penalizzando in modo mirato le aree che risultano meno realistiche e contribuendo così a generare immagini globalmente più coerenti e naturali.

Nei paragrafi successivi verranno descritte in dettaglio da un punto di vista teorico tutte le principali componenti alla base delle architetture del generatore e del discriminatore.

Convolutional Layers

I *convolutional layers* rappresentano il nucleo delle Convolutional Neural Networks (CNNs), in quanto responsabili dell'estrazione automatica delle caratteristiche discriminanti dalle immagini. A differenza delle reti neurali tradizionali (*fully-connected*), in cui ogni neurone è connesso a tutti i pixel dell'input, nei convolutional layers ciascun neurone è connesso soltanto a una regione locale dell'immagine, detta *receptive field*. Questa caratteristica consente di ridurre drasticamente il numero di parametri, migliorando l'efficienza computazionale e diminuendo il rischio di *overfitting*, uno dei principali limiti delle architetture dense [35].

Il principio operativo alla base di questi strati è la *convoluzione*. Sia I l'immagine di input e K un filtro (o kernel), l'operazione di convoluzione bidimensionale può essere formalizzata come:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n)\, K(m,n),$$

dove $S(i,j)$ rappresenta l'elemento (i,j) della *feature map* risultante. Attraverso questa operazione, il filtro viene traslato lungo l'immagine eseguendo prodotti scalari tra i pesi di K e le regioni locali di I. Ogni filtro è in grado di estrarre particolari caratteristiche, agendo come "rilevatore di patterns" (es. bordi, angoli o texture), e l'impiego di più filtri consente di costruire rappresentazioni gerarchiche, in cui gli strati iniziali catturano caratteristiche di basso livello, mentre quelli più profondi estraggono concetti via via più astratti [36]. I valori dei filtri (pesi) non sono fissati a priori, ma vengono appresi durante il training mediante backpropagation, minimizzando l'errore della rete.

Il risultato dell'operazione di convoluzione è una *feature map* (o *activation map*), che rappresenta le informazioni più rilevanti dell'immagine, come bordi, texture e forme locali. Rispetto all'immagine originale, queste mappe contengono meno ridondanza e conservano solo le strutture discriminanti utili per l'apprendimento.

L'output di un livello convoluzionale dipende da tre iperparametri fondamentali:

- **Profondità** (*depth*) — indica il numero di filtri applicati e, di conseguenza, il numero di *feature map* prodotte. Ogni filtro apprende un diverso tipo di caratteristica visiva (ad esempio bordi, texture o gradienti di colore).
- **Stride** (S) — rappresenta il passo con cui il filtro si sposta sull'immagine. Valori maggiori di S riducono la dimensione spaziale dell'output, mentre valori più piccoli permettono una rappresentazione più dettagliata.
- **Zero-padding** (Z) — consiste nell'aggiungere zeri ai bordi dell'input per preservarne la dimensione e prevenire la perdita di informazioni ai margini.

La dimensione spaziale dell'output O di un livello convoluzionale, dato un input di dimensione V, un filtro di dimensione R, stride S e padding Z, è calcolata come:

$$O = \frac{V - R + 2Z}{S} + 1.$$

Un principio chiave delle reti convoluzionali è il *parameter sharing*. Esso si basa sull'assunzione che una stessa caratteristica (ad esempio un bordo orizzontale o una curva) possa comparire in posizioni diverse dell'immagine. Per questo motivo, i pesi di ciascun filtro vengono condivisi tra tutte le posizioni spaziali: ciò riduce drasticamente il numero di parametri da apprendere, semplificando il modello senza comprometterne la capacità di generalizzazione [35].

Residual Blocks

I *residual blocks* rappresentano una delle innovazioni più significative nello sviluppo delle reti neurali profonde. Grazie all'introduzione delle *skip connections*, le Residual Network (ResNet) hanno permesso di addestrare modelli con profondità senza precedenti, migliorando le prestazioni nei compiti di classificazione e riconoscimento di oggetti e stabilendo un nuovo standard nello *state-of-the-art* delle architetture convoluzionali [37, 38].

Con l'aumentare della profondità delle reti, è emerso un problema noto come *degradation problem*: aggiungere ulteriori layer non porta necessariamente a una migliore accuratezza e può addirittura peggiorare le prestazioni sul training set. Questo fenomeno non è dovuto a *overfitting*, ma a difficoltà di ottimizzazione che impediscono alla rete di apprendere correttamente funzioni complesse.
Per risolvere tale criticità, He et al. (2016) [37] hanno introdotto il paradigma del *residual learning*. L'idea di base è semplice: invece di apprendere direttamente una funzione obiettivo $H(x)$, si addestra la rete ad apprendere la funzione residua $F(x)$ che misura la differenza tra l'output desiderato e l'input originale:

$$F(x) = H(x) - x \quad \Longrightarrow \quad H(x) = F(x) + x.$$

In questo modo, la rete si concentra sull'apprendimento delle variazioni rispetto all'identità: se la funzione identità stessa rappresentasse la soluzione ottimale, sarebbe sufficiente che la rete imparasse $F(x) \approx 0$, riducendo così la complessità del processo di ottimizzazione.
L'introduzione di questo approccio ha portato due vantaggi fondamentali: da un lato, ha eliminato le difficoltà di ottimizzazione legate alla profondità e prevenuto la degradazione delle prestazioni; dall'altro, ha reso possibile la costruzione di architetture estremamente profonde (fino a 152 layer su ImageNet) mantenendo stabilità e efficienza computazionale, superando modelli precedenti come VGG e GoogLeNet [37].

Tale concetto viene implementato tramite le *connessioni residue* (*skip connections*), che consentono di sommare direttamente l'input x all'output trasformato $F(x)$:

$$y = F(x, \{W_i\}) + x,$$

dove $F(x, \{W_i\})$ è una sequenza di operazioni (convoluzioni, normalizzazioni e attivazioni non lineari) parametrizzata dai pesi W_i. A seconda delle necessità architetturali, le connessioni possono essere:

- **Identity shortcuts**, che copiano direttamente l'input senza introdurre nuovi parametri;
- **Projection shortcuts**, che impiegano convoluzioni 1×1 per adattare dimensioni differenti tra input e output.

Dal punto di vista implementativo, si distinguono due principali varianti:

- **Basic block** — utilizzato nelle versioni meno profonde (ResNet-18 e ResNet-34), composto da due convoluzioni 3×3;
- **Bottleneck block** — adottato nelle versioni più profonde (ResNet-50, ResNet-101, ResNet-152), che combina una sequenza 1×1 – 3×3 – 1×1, dove le convoluzioni 1×1 riducono e successivamente ripristinano la dimensionalità per migliorare l'efficienza.

L'integrazione dei residual blocks all'interno del generatore della CycleGAN consente di ottenere una rappresentazione stabile e coerente, capace di apprendere trasformazioni visive complesse mantenendo la qualità strutturale delle immagini di input.

Instance Normalization

La *Instance Normalization (IN)*, introdotta da Ulyanov et al. [39], è una tecnica di normalizzazione che ha trovato ampio impiego nelle reti generative, in particolare nei compiti di *style transfer* e generazione di immagini. L'idea alla base è simile alla Batch Normalization (BN), ma con una differenza cruciale: mentre la BN normalizza le attivazioni calcolando media e varianza sull'intero batch, la IN esegue la normalizzazione per singola immagine e per canale [40].

Formalmente, dato un tensore di input $x \in \mathbb{R}^{T \times C \times H \times W}$ (batch di T immagini con C canali e dimensioni spaziali $H \times W$), l'operazione di instance normalization per il t-esimo esempio e l'i-esimo canale è definita come:

$$y_{tijk} = \frac{x_{tijk} - \mu_{ti}}{\sqrt{\sigma_{ti}^2 + \epsilon}},$$

dove

$$\mu_{ti} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H} x_{tilm}, \quad \sigma_{ti}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{tilm} - \mu_{ti})^2.$$

Qui μ_{ti} e σ_{ti}^2 rappresentano rispettivamente la media e la varianza calcolate per ciascuna immagine e ciascun canale, mentre ϵ è un termine di stabilizzazione numerica.

Secondo quanto riportato da [40], l'Instance Normalization (IN) può essere facilmente integrata in diverse architetture di reti neurali, come le CNNs e le recurrent neural networks (RNNs). Nelle CNN, ad esempio, l'IN può essere applicata dopo ogni layer convoluzionale, contribuendo a ottenere invarianza alla traslazione, migliorando la stabilità e la capacità di generalizzazione del modello.
Inoltre, l'IN è stata ampiamente impiegata anche nelle GAN per incrementare la qualità delle immagini sintetiche: normalizzando le feature a livello di singola istanza, infatti, essa elimina la dipendenza dalla media e dalla varianza del mini-batch, rendendo la rete meno sensibile a variazioni nella distribuzione dei dati.

Un ulteriore vantaggio risiede nell'interpretabilità delle reti: l'IN produce rappresentazioni delle features più coerenti con le informazioni semantiche, aspetto cruciale in applicazioni come l'image-to-image translation. Inoltre, rispetto ad altre tecniche di normalizzazione come la batch normalization, l'IN presenta benefici computazionali, in quanto può essere applicata a singole istanze in parallelo all'interno di un batch, con conseguente riduzione dei tempi di calcolo.
Complessivamente, l'adattabilità dell'Instance Normalization a differenti architetture evidenzia la sua versatilità e il potenziale nel migliorare un'ampia gamma di compiti nel Deep Learning.

Funzioni di attivazione ReLU e Leaky ReLU

La *Rectified Linear Unit (ReLU)* è una delle funzioni di attivazione più utilizzate nelle reti neurali profonde per via della sua semplicità computazionale e della capacità di mitigare il problema del gradiente che svanisce, tipico di funzioni sigmoidi e tangenti iperboliche. Definita come

$$f(x) = \max(0, x),$$

la ReLU attiva il neurone soltanto per valori positivi, imponendo una soglia a zero. Questa caratteristica accelera il processo di apprendimento e consente una migliore propagazione del gradiente nei livelli profondi [41]. Tuttavia, un limite noto è il cosiddetto *dying ReLU problem*, in cui i neuroni possono rimanere permanentemente inattivi (cioè produrre solo output nulli) se i pesi portano l'input a valori negativi [42].

Per affrontare questa criticità, è stata proposta la *Leaky ReLU*, una variante della ReLU che introduce un piccolo coefficiente $\alpha > 0$ per i valori negativi:

$$f(x) = \begin{cases} x, & \text{se } x \geq 0, \\ \alpha x, & \text{se } x < 0. \end{cases}$$

In questo modo, invece di annullare completamente le attivazioni negative, si mantiene un gradiente diverso da zero, riducendo il rischio che i neuroni "muoiano" e migliorando la capacità di apprendimento della rete [43]. La Leaky ReLU ha dimostrato empiricamente di favorire la convergenza in contesti in cui la ReLU standard tende a bloccare il flusso di informazioni.

Il discriminatore PatchGAN

Il discriminatore PatchGAN è un componente introdotto da Isola et al. (2017) [44] nel modello *Pix2Pix* e successivamente ripreso in architetture come *CycleGAN* [22]. A differenza dei discriminatori classici che valutano un'intera immagine come autentica o generata, il PatchGAN discrimina a livello locale, suddividendo l'immagine in sotto-regioni ("patch") e classificando ciascuna di esse come reale o falsa. L'output finale è una mappa bidimensionale di probabilità, in cui ogni cella corrisponde al giudizio del discriminatore su una patch della dimensione predefinita.

L'obiettivo del PatchGAN è modellare l'immagine come una distribuzione di patch locali, ipotizzando che la qualità globale di un'immagine sintetica dipenda dalla sua coerenza locale. In pratica, si applica una rete convoluzionale che scorre sull'immagine con finestra $N \times N$ (tipicamente 70×70 pixel), producendo una decisione di real/fake per ciascuna regione:

$$D(x) \in \mathbb{R}^{H' \times W'},$$

dove H' e W' dipendono dalla dimensione dell'input e dalla dimensione della patch. Ogni elemento di $D(x)$ rappresenta la valutazione di una patch.

Il discriminatore PatchGAN può essere visto come una CNN che riduce progressivamente le dimensioni spaziali fino ad arrivare a una griglia di uscite. Ogni convoluzione è tipicamente seguita da una funzione di attivazione *Leaky ReLU* [45]. Il vantaggio del PatchGAN risiede nel fatto che, concentrandosi su piccole regioni, il discriminatore apprende le statistiche locali di texture, bordi e dettagli, elementi fondamentali per garantire realismo visivo. Inoltre, rispetto a un discriminatore che valuta l'intera immagine, riduce il numero di parametri e accelera l'addestramento. Nelle architetture *image-to-image translation*, come Pix2Pix e CycleGAN, l'impiego del PatchGAN si è dimostrato particolarmente efficace perché preserva la coerenza locale tra immagine di input e immagine generata, guida il generatore ad apprendere texture realistiche e riduce il rischio di generare immagini sfocate, problema tipico di altre formulazioni di loss.

2.3 Addestramento del modello

In questa sezione vengono descritti nel dettaglio tutti gli aspetti relativi all'*addestramento* della CycleGAN per fornire una descrizione completa dei parametri e delle opzioni di training.
Per la realizzazione del modello è stata impiegata la stessa configurazione della rete proposta nel lavoro originale di Zhu et al. (2017) [22], poiché tale architettura si è dimostrata efficace in numerosi contesti di generazione di immagini. Questa stessa configurazione è stata successivamente ripresa anche da Jütte et al. (2024) [1] nei suoi esperimenti, confermandone la robustezza e la capacità di adattarsi allo scenario applicativo oggetto di questo studio. La decisione di mantenere questa architettura standard deriva dunque dalla volontà di basarsi su un modello consolidato e validato in letteratura, così da concentrare l'attenzione dell'analisi sulle prestazioni e sulla qualità delle immagini generate, piuttosto che sulla progettazione di una nuova rete.
Infine, saranno illustrate le *misure di performance* utilizzate per monitorare l'andamento dell'apprendimento e valutare, per quanto possibile, la correttezza del processo di ottimizzazione: questo tipo di analisi può consentire di identificare se l'addestramento è caratterizzato da una convergenza stabile e da un equilibrio tra i due modelli avversari. Equivalentemente, da ora in avanti, potremo riferirci a tali misure di performance con il termine "*metriche*", poiché tipicamente usato nella comunità informatica, da non confondere con il concetto matematico di metrica (cioè distanza).

2.3.1 Implementazione e scelta dei parametri

Il modello CycleGAN, come descritto da Zhu et al. (2017) [22], impiega due generatori e due discriminatori che operano in modo avversario per trasformare le immagini tra due domini, che nel nostro caso sono rappresentati dalle immagini di nevo benigno e melanoma. L'addestramento è stato condotto seguendo la configurazione proposta dagli autori, opportunamente adattata al contesto sperimentale di questo lavoro. L'utilizzo di un'architettura già validata e ampiamente riconosciuta in letteratura ha permesso di evitare la necessità di sviluppare un nuovo modello da zero, concentrando invece gli sforzi sull'analisi e sull'ottimizzazione del processo generativo. Questa scelta ha garantito un utilizzo efficiente delle risorse computazionali, permettendo al tempo stesso di raggiungere l'obiettivo di generare immagini dermoscopiche sintetiche realistiche.
Tutti gli esperimenti sono stati condotti in *Python* su una NVIDIA RTX 500 Ada Generation Laptop GPU, che ha garantito tempi di calcolo sufficientemente compatibili con le esigenze sperimentali.

Dataset e preprocessing Il dataset personalizzato è stato costruito unendo due directory, contenenti rispettivamente immagini di nevi e melanomi. Le immagini sono state preprocessate utilizzando la libreria *Albumentations*, con le seguenti trasformazioni:

- Ridimensionamento a 286×286 pixel.
- *RandomCrop* a 256×256 per introdurre variabilità spaziale.
- Normalizzazione con media $[0.5, 0.5, 0.5]$ e deviazione standard $[0.5, 0.5, 0.5]$.
- Conversione in tensori tramite *ToTensorV2()*.

Il dataset è stato suddiviso in due insiemi: "*training*" e "*validation*", con i dati caricati tramite "*DataLoader*" di PyTorch, utilizzando un *batch size = 1* e con *shuffle attivo* sul training set.

Generatori e discriminatori: Il *generatore* segue una struttura modulare di tipo *encoder–transformer–decoder*, composta da due blocchi principali:

- **Blocco convoluzionale con downsampling o upsampling opzionale**: Il generatore utilizza due tipi di blocchi convoluzionali, uno per il downsampling e l'altro per l'upsampling, che sfrutta una convoluzione trasposta. I blocchi di downsampling e upsampling sono costituiti da una convoluzione seguita da *InstanceNorm* per stabilizzare l'apprendimento e ReLU per introdurre non-linearità, implementati con la libreria *torch.nn*.
- **Blocco residuo**: Il blocco residuo apprende la mappatura residua tra l'input e l'output. Il primo blocco convoluzionale della sequenza include una funzione di attivazione, che aiuta a catturare e estrarre le caratteristiche importanti dall'input, mentre il secondo blocco si concentra principalmente sull'adattamento delle dimensioni delle caratteristiche estratte, senza includere una funzione di attivazione. Le informazioni residue vengono quindi sommate direttamente all'input originale.

Il nostro generatore è composto da due strati di downsampling, seguiti da nove blocchi residui, e infine da due strati di upsampling e la sua struttura è descritta come segue:

- Un blocco convoluzionale iniziale (*Convolution–InstanceNorm–ReLU*)con kernel 7×7, 64 filtri e stride 1.
- Due blocchi convoluzionali con downsampling, con kernel 3×3 e stride 2, rispettivamente con 128 e 256 filtri.

- Nove blocchi residui, ognuno con due convoluzioni 3×3, padding 1, e 256 filtri.
- Due strati di upsampling, con stride frazionario 1/2 e 128 e 64 filtri.
- Un blocco convoluzionale finale (*Convolution–InstanceNorm–ReLU*) con kernel 7×7, 3 filtri e stride 1.

Questa architettura, validata in letteratura, ha dimostrato di essere efficace nella generazione di immagini realistiche, facilitando la trasformazione bidirezionale tra lesioni benigne e maligne, e si è adattata adeguatamente alle esigenze di questo studio; ne viene mostrata una rappresentazione in Figura 2.10.

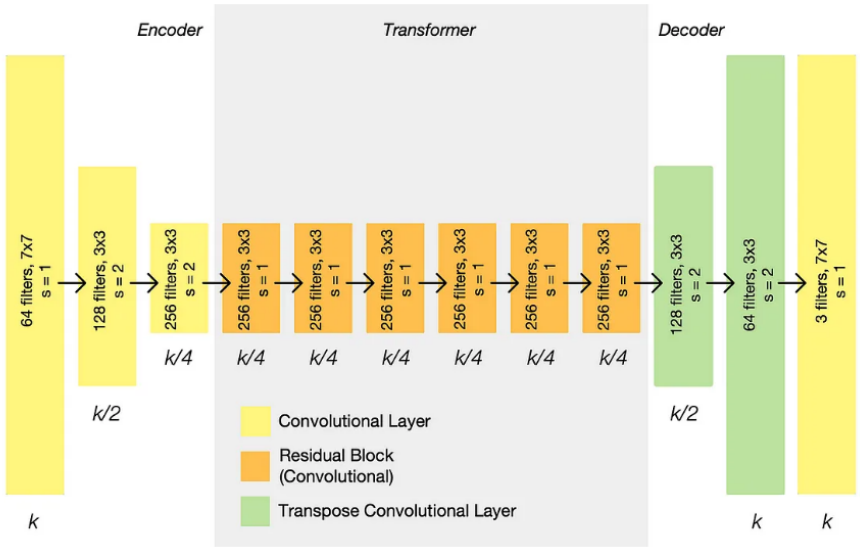


Figura 2.10: Architettura del generatore con 6 residual blocks [46]. Nel presente lavoro sono stati invece impiegati *9 residual blocks*, in linea con quanto riportato da Zhu et al. (2017) [22] per immagini di dimensione (256 x 256).

Il *discriminatore* adotta l'approccio *PatchGAN* 70x70, progettato per valutare il realismo delle immagini a livello locale, anziché globale.
L'architettura del discriminatore è definita come una sequenza di *Convolution –InstanceNorm -LeakyReLU* con kernel 4×4, stride 2 e k filtri, con $k = 64, 128, 256, 512$. La funzione di attivazione *LeakyReLU* ha un coefficiente $\alpha = 0.2$, mentre l'ultimo layer convoluzionale produce una mappa di probabilità 1-dimensionale, in cui ogni valore rappresenta la stima di realismo di una specifica patch dell'immagine in input. Questo approccio consente al discriminatore di concentrarsi sui dettagli locali (bordo, texture, variazioni cromatiche), migliorando la capacità del modello di distinguere le immagini sintetiche da quelle reali.

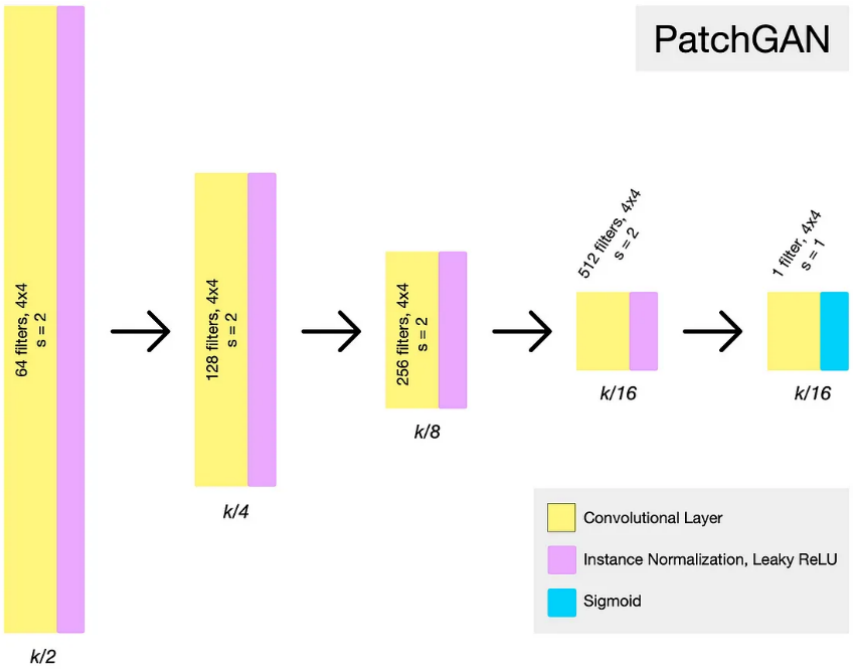


Figura 2.11: Architettura del discriminatore PatchGAN [46].

Loss function: Nel presente lavoro, come nella rete proposta da Zhu et al. (2017) [22], la rete generativa adotta la formulazione della *Least Squares Generative Adversarial Network (LSGAN)*, introdotta da Mao et al. (2017) [47]. Questa variante nasce con l'obiettivo di rendere l'addestramento più stabile e migliorare la qualità visiva delle immagini, superando uno dei limiti principali delle GAN classiche: il *vanishing gradient problem*, ovvero la perdita di segnale di errore nel generatore quando il discriminatore diventa eccessivamente accurato.

Nella GAN originale, il discriminatore è addestrato come un classificatore binario con una funzione di perdita basata sulla *sigmoid cross-entropy*. Tuttavia, quando il generatore produce immagini già collocate sul lato "corretto" della frontiera di decisione (cioè riconosciute come false), la loss genera un gradiente quasi nullo: il generatore smette di migliorare pur non avendo ancora raggiunto un livello di realismo soddisfacente. Per risolvere questo problema, la LSGAN sostituisce la *cross-entropy loss* con una loss ai minimi quadrati, che penalizza proporzionalmente la distanza tra l'output del discriminatore e il valore target desiderato. In questo modo, anche campioni sintetici già correttamente classificati ma ancora lontani dal dominio reale contribuiscono al gradiente, mantenendo un segnale di apprendimento informativo per il generatore.

La formulazione adottata è la seguente:

$$\min_D \mathcal{L}_{\text{LSGAN}}(D) = \frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{data}}(x)}[(D(x)-1)^2] + \frac{1}{2}\,\mathbb{E}_{z\sim p_z(z)}[(D(G(z)))^2],$$
$$\min_G \mathcal{L}_{\text{LSGAN}}(G) = \frac{1}{2}\,\mathbb{E}_{z\sim p_z(z)}[(D(G(z))-1)^2].$$

Rispetto alla formulazione standard di Goodfellow et al. (2014) [21], la LSGAN presenta diversi vantaggi [47]:

- **Gradienti più stabili:** la penalizzazione quadratica evita l'annullamento del gradiente quando il discriminatore è troppo preciso, mantenendo un segnale di apprendimento costante per il generatore;
- **Immagini più realistiche:** la rete è incoraggiata a produrre campioni vicini alla frontiera decisionale, riducendo la distanza percettiva dal dominio reale;
- **Addestramento più robusto:** la loss ai minimi quadrati attenua le oscillazioni tipiche delle GAN classiche e riduce il rischio di *mode collapse*.

Per l'addestramento, quindi, le componenti della loss function implementate come segue:

- **Adversarial loss** — serve a far sì che i generatori $G_{X\to Y}$ e $G_{Y\to X}$ producano immagini realistiche nei due domini X (nevi) e Y (melanomi), mentre i discriminatori D_X e D_Y imparano a distinguerle dalle immagini reali. È stata implementata tramite *Mean Squared Error (MSE)* secondo la configurazione LSGAN .

  Le loss dei due generatori vengono calcolate separatamente e sommate:

$$\mathcal{L}_{\text{adv}}^{G_{X\to Y}} = \frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[(D_Y(G_{X\to Y}(x))-1)^2\Big]$$
$$\mathcal{L}_{\text{adv}}^{G_{Y\to X}} = \frac{1}{2}\,\mathbb{E}_{y\sim p_{\text{data}}(y)}\Big[(D_X(G_{Y\to X}(y))-1)^2\Big]$$

  Ciascun termine misura quanto l'immagine sintetica generata si discosti dall'essere riconosciuta come reale dal discriminatore corrispondente.
  La loss adversarial complessiva per i due generatori è quindi data da:

$$\mathcal{L}_{\text{adv}}^{G} = \mathcal{L}_{\text{adv}}^{G_{X\to Y}} + \mathcal{L}_{\text{adv}}^{G_{Y\to X}}.$$

  Per i discriminatori si ha invece:

$$\mathcal{L}_{\text{adv}}^{D_Y} = \frac{1}{2}\,\mathbb{E}_{y\sim p_{\text{data}}(y)}\Big[(D_Y(y)-1)^2\Big] + \frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[D_Y(G_{X\to Y}(x))^2\Big]$$
$$\mathcal{L}_{\text{adv}}^{D_X} = \frac{1}{2}\,\mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[(D_X(x)-1)^2\Big] + \frac{1}{2}\,\mathbb{E}_{y\sim p_{\text{data}}(y)}\Big[D_X(G_{Y\to X}(y))^2\Big]$$

$$\mathcal{L}_{\text{adv}}^{D} = \mathcal{L}_{\text{adv}}^{D_Y} + \mathcal{L}_{\text{adv}}^{D_X}$$

Il principio di base è che il discriminatore cerca di assegnare il valore 1 alle immagini reali e 0 a quelle sintetiche, mentre il generatore cerca di produrre immagini che il discriminatore classifichi con un valore vicino a 1.

- **Cycle-consistency loss** - termine che garantisce la coerenza strutturale tra domini e la conservazione delle caratteristiche morfologiche, con peso λ_{cyc}. È calcolata come la distanza L_1 tra l'immagine originale e quella ricostruita dopo la doppia trasformazione $X \to Y \to X$ e $Y \to X \to Y$.

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[\|G_{Y\to X}(G_{X\to Y}(x)) - x\|_1\Big] \tag{2.2}$$
$$+\,\mathbb{E}_{y\sim p_{\text{data}}(y)}\Big[\|G_{X\to Y}(G_{Y\to X}(y)) - y\|_1\Big] \tag{2.3}$$

- **Identity loss** - termine opzionale che mira a preservare i colori originali nelle immagini generate, imponendo che un'immagine già nello stile di destinazione, se passata come input a un generatore, venga idealmente restituita invariata. È calcolata come la distanza media in norma L_1 tra un'immagine reale e la stessa immagine dopo essere stata passata nel generatore "sbagliato":

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{x\sim p_{\text{data}}(x)}\Big[\|G_{Y\to X}(x) - x\|_1\Big] + \mathbb{E}_{y\sim p_{\text{data}}(y)}\Big[\|G_{X\to Y}(y) - y\|_1\Big].$$

La funzione di perdita complessiva del modello è data dalla somma pesata dei tre termini:

$$\mathcal{L} = \mathcal{L}_{\text{adv}}^{G} + \mathcal{L}_{\text{adv}}^{D} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}},$$

dove i coefficienti λ_{cyc} e λ_{id} controllano rispettivamente l'importanza del vincolo di coerenza ciclica e della fedeltà cromatica. In questo lavoro, i valori di tali pesi sono stati scelti in accordo con la letteratura, con λ_{cyc} impostato a 10 e λ_{id} a 0.

Strategia di training: Per entrambi i network è stato utilizzato l'*ottimizzatore Adam*, con un learning rate iniziale di 2×10^{-4} e parametri $(\beta_1, \beta_2) = (0.5, 0.999)$. L'addestramento è stato eseguito per un totale di *200 epoche*, con *batch size* pari a 1, come suggerito dalla letteratura, per garantire un buon equilibrio tra qualità delle immagini e tempi di addestramento compatibili con le risorse disponibili. È stato applicato un decadimento lineare del learning rate a partire dalla centesima epoca, per favorire la convergenza e ridurre le oscillazioni nelle epoche finali, secondo la formula:

$$\lambda(e) = \begin{cases} 1 & \text{se } e < 100, \\ 1 - \dfrac{e - 100}{N_{\text{epoche}} - 100} & \text{altrimenti.} \end{cases}$$

Questa strategia ha permesso una discesa più morbida e un miglior equilibrio nella fase finale dell'addestramento.

Per ridurre le oscillazioni del modello durante il training, come fatto da [22], i discriminatori sono stati aggiornati utilizzando una memoria storica di immagini generate, anziché limitarsi a quelle prodotte dall'iterazione più recente dei generatori. Invece di aggiornare i discriminatori solo con le immagini generate nell'ultimo passo del training, si utilizza una *memoria temporanea* (buffer) che contiene le ultime 50 immagini sintetiche prodotte dai generatori. Durante l'addestramento, a ogni passo vengono selezionate casualmente alcune immagini da questo buffer per l'aggiornamento dei discriminatori, mentre le nuove immagini generate sostituiscono progressivamente quelle più vecchie. Questo meccanismo evita che i discriminatori si adattino troppo rapidamente alle variazioni dei generatori e contribuisce a mantenere più stabile il processo di apprendimento.

È stata inoltre utilizzata la tecnica di *addestramento a precisione mista* (*mixed precision training*) tramite "*torch.amp.GradScaler()*", che gestisce automaticamente lo scaling dei gradienti. Ogni epoca è stata completata con il salvataggio di immagini di esempio (*input*, *fake melanoma*, *fake nevus*) per monitorare visivamente i progressi del modello e fare un confronto diretto fra immagine di partenza reale e generata nel nuovo dominio. I principali parametri e i loro corrispondenti valori sono riportati e riassunti nella Tabella 2.3.

2.3.2 Analisi delle metriche di addestramento

Durante l'addestramento di una rete neurale, è fondamentale monitorare l'evoluzione delle metriche al crescere delle epoche per verificare che il training stia procedendo correttamente. Nel caso delle GAN, tale analisi risulta tuttavia più complessa a causa della natura competitiva del problema di ottimizzazione: la dinamica di apprendimento tra generatore e discriminatore può infatti generare fenomeni di

Parametro	Valore
Dimensioni immagini	256x256
Rete generatore	ResNet - 9 blocchi
Rete discriminatore	PatchGAN
Loss avversaria	MSE (LSGAN)
Loss ciclica	distanza L_1
λ_{cyc}	10
λ_{id}	0
Epoche	200
Batch size	1
Ottimizzatore	Adam
β_1	0.5
β_2	0.999
LR iniziale	2×10^{-4}
LR scheduler	Lineare, da epoca 100

Tabella 2.3: Parametri di addestramento del modello CycleGAN.

instabilità o divergenza, rendendo necessario un controllo costante delle metriche per garantire una convergenza bilanciata tra le due reti. Allo stesso tempo, è importante interpretare tali valori in combinazione con altre metriche di valutazione, interne ed esterne, come quelle che verranno descritte nelle sezioni successive. In particolare, nel presente lavoro sono stati analizzati due indicatori principali per quanto riguarda l'analisi del processo di addestramento: la *training loss* e l'*output medio dei discriminatori*. Di seguito viene fornita una descrizione teorica del loro andamento atteso, secondo quanto riportato in letteratura.

Training loss: L'andamento delle *loss function* rappresenta un indicatore diretto della stabilità del processo di ottimizzazione e del livello di equilibrio raggiunto tra le due parti della rete. Nel caso delle GAN, l'analisi delle loss è più delicata rispetto ai modelli tradizionali, poiché il loro andamento non riflette necessariamente un miglioramento monotono: si tratta infatti di un gioco competitivo min–max in cui ogni rete cerca di ottimizzare il proprio obiettivo in contrasto con l'altra.
Per questo motivo, per valutare il corretto andamento del training è essenziale monitorare separatamente le diverse componenti della loss: quella avversaria del generatore, quella del discriminatore e la cycle-consistency loss, le cui formulazioni sono state discusse nell'implementazione.

In condizioni ideali, come illustrato da Sambath et al. (2022) [48], la *loss complessiva del generatore* parte da valori elevati e decresce rapidamente nelle prime epoche, stabilizzandosi successivamente su livelli più bassi e regolari.

La *loss del discriminatore*, invece, diminuisce in modo più graduale e tende ad assestarsi su un valore intermedio stabile, senza oscillazioni marcate, indicando così che il discriminatore continua a distinguere efficacemente le immagini reali da quelle sintetiche, ma senza prevalere sul generatore. Una dinamica di questo tipo riflette un buon equilibrio del gioco min–max, in cui nessuna delle due reti domina sull'altra.
Infine, la *cycle-consistency loss* dovrebbe decrescere in modo monotono fino a stabilizzarsi su una soglia costante, indicando che la rete ha appreso una mappatura coerente tra i due domini, senza collassare verso una trasformazione identitaria.

L'andamento empirico delle loss osservato durante il training verrà presentato e discusso nel Capitolo 3, dove si mostrerà come le curve ottenute riflettano il comportamento atteso, confermando la stabilità complessiva del processo di addestramento.

Output medio dei discriminatori: Oltre alle principali funzioni di loss, sono stati monitorati anche i valori medi degli output dei discriminatori. In particolare, abbiamo analizzato quello relativo al dominio delle immagini di melanoma, per avere una misura diretta del suo comportamento. Questi valori sono stati indicati come:

- "$M_{\mathbf{reals}}$": la media dell'output del discriminatore quando vengono fornite immagini reali di melanoma;
- "$M_{\mathbf{fakes}}$": la media dell'output dello stesso discriminatore quando riceve immagini sintetiche generate dal corrispondente generatore. "M_{reals}"

Poiché la loss del discriminatore è calcolata come *Mean Squared Error (MSE)* rispetto ai target 1 per le immagini reali e 0 per quelle sintetiche, i valori medi di "M_{reals}" e "M_{fakes}" riflettono la sicurezza con cui il discriminatore distingue le due categorie:

- valori di "M_{reals}" prossimi a 1 indicano un'elevata confidenza nel riconoscere le immagini reali;
- valori di "M_{fakes}" vicini a 0 segnalano una corretta identificazione delle immagini sintetiche come false.

Secondo quanto riportato da [47], nelle primissime iterazioni del training con configurazione LSGAN, il discriminatore tende a migliorare rapidamente, poiché le immagini sintetiche prodotte dal generatore sono visivamente molto lontane dal dominio reale, imparando a riconoscere facilmente i campioni reali (valori di "M_{real}" elevati) e a respingere quelli sintetici (valori di "M_{fake}" bassi). Il generatore, invece,

nelle prime epoche fatica maggiormente, poiché deve imparare a modellare una distribuzione complessa a partire da input appartenenti a un dominio diverso.
Con il progredire dell'addestramento, man mano che le immagini sintetiche diventano più realistiche, il discriminatore riduce la propria sicurezza nelle predizioni: "M_{fakes}" tende gradualmente ad aumentare, mentre "M_{reals}" può leggermente diminuire. Questo comportamento segnala un miglioramento qualitativo delle immagini e un equilibrio crescente tra le due reti.

Idealmente, i due valori dovrebbero convergere verso un intervallo intermedio (intorno a 0.5), condizione che rappresenta un *equilibrio avversario* approssimato o equilibrio di Nash: il discriminatore non è più in grado di distinguere con certezza le immagini reali da quelle sintetiche, mentre il generatore ha appreso a produrre campioni visivamente credibili. Nella pratica, tuttavia, la letteratura non riporta evidenze che gli output dei discriminatori in LSGAN o CycleGAN convergano esattamente a 0.5. La funzione ai minimi quadrati tende infatti a spingere i valori verso 0 o 1, mantenendo un margine di separazione anche in condizioni di equilibrio. Inoltre, l'addestramento avversario rimane intrinsecamente dinamico: le due reti si aggiornano in modo alternato, modificando continuamente le rispettive distribuzioni e impedendo una convergenza perfetta.
Pertanto, sulla base del funzionamento di questa tipologia di rete e delle informazioni raccolte dalla letteratura, ipotizziamo che valori medi oscillanti in intervalli intermedi, ad esempio "M_{reals}" $\approx 0.6 - 0.7$ e "M_{fakes}" $\approx 0.3 - 0.4$, possano essere considerati indicativi di un equilibrio realistico e di un training ben bilanciato, anche se non perfettamente simmetrico. In generale, ci aspetteremmo che la differenza tra "M_{reals}" e "M_{fakes}" si riduca progressivamente senza annullarsi del tutto, a indicare che le due reti hanno raggiunto una competizione stabile.
Al contrario, se il discriminatore prevale sul generatore, "M_{reals}" dovrebbe rimanere vicino a 1 e "M_{fakes}" vicino a 0 anche dopo molte epoche, ostacolandone il miglioramento; viceversa, concluderemmo che il discriminatore risulta inefficace se i due valori convergono troppo rapidamente verso 0.5: in tal caso il sistema può incorrere in fenomeni di *mode collapse*, con immagini ripetitive o incoerenti.

È importante precisare che, sebbene quanto descritto derivi dallo studio approfondito dell'architettura implementata e dal confronto con le analisi riportate in letteratura, si tratta di ipotesi non confermate, ma comunque motivate. È anche opportuno ricordare che nei lavori originali la convergenza del training delle GAN viene generalmente caratterizzata attraverso altre modalità, come l'andamento delle loss function o l'approssimazione dell'estimation error bound. In questo contesto, l'analisi dell'output medio dei discriminatori potrebbe rappresentare un contributo innovativo, offrendo una prospettiva empirica alternativa per valutare il comportamento dinamico del modello e la stabilità del processo di apprendimento, potenzialmente rilevante per studi futuri.

2.4 Validazione

Una volta completato l'addestramento e ottenute le immagini dermoscopiche sintetiche, l'impressione visiva iniziale era decisamente positiva: le lesioni generate sembravano riprodurre in modo convincente le caratteristiche tipiche dei nevi e dei melanomi, rispettando, almeno qualitativamente e a prima vista, i criteri della regola ABCDE. Terminata la presentazione dei primi risultati, tuttavia, la domanda è sorta spontanea: *"Chi ci assicura davvero che queste immagini siano realistiche e rappresentino in modo corretto le caratteristiche peculiari delle diverse tipologie di lesione, se non un dermatologo?"*
Non disponendo di una valutazione clinica esperta, l'obiettivo successivo è stato quello di esplorare strategie alternative per sostituire idealmente il giudizio del dermatologo, cercando metodi di valutazione oggettivi, quantitativi e riproducibili, che andassero oltre l'impressione visiva (il semplice "sembrano buone").

In letteratura, la qualità delle immagini generate dalle GAN viene spesso valutata attraverso metriche che restituiscono un valore numerico associato a due aspetti principali: la *fedeltà* e la *varietà*.
Alcune metriche combinano entrambe le componenti in un singolo valore, come il *Fréchet Inception Distance (FID)* [49] e il *Kernel Inception Distance (KID)* [50]; altre, come la metrica *Precision, Recall, Density, Coverage (PRDC)* [51], distinguono esplicitamente le due dimensioni, consentendo un'analisi più dettagliata. Metriche come l'Inception Score (IS) [52], invece, sono state inizialmente considerate, ma poi scartate, poiché risultano meno significative nel nostro contesto applicativo, come evidenziato in alcuni articoli scientifici sul tema [53].

Parallelamente, per cercare di rispondere al quesito iniziale, abbiamo formulato la seguente osservazione: *"Se avessimo a disposizione un buon classificatore, addestrato su immagini reali di lesioni cutanee, capace di distinguere tra melanomi e nevi benigni, potremmo usarlo per valutare la fedeltà delle immagini generate!"* L'idea è che, se le immagini sintetiche risultano sufficientemente realistiche, un classificatore accurato dovrebbe riconoscerle correttamente, attribuendo loro probabilità coerenti con la loro classe di appartenenza. Chiaramente, alla base di questa intuizione si presuppone che tale classificatore sia in grado di riconoscere e distinguere efficacemente le features che contraddistinguono le due classi di lesioni, aspetto da non sottovalutare affatto.

Dalle considerazioni sopra menzionate è nata e si è sviluppata la nostra strategia di validazione, che combina l'analisi delle metriche di similarità più diffuse con la valutazione tramite classificatori esterni, scelti tra modelli già validati in letteratura e per i quali sono disponibili i pesi pre-addestrati. Questo approccio ci ha permesso di evitare ulteriori fasi di training lunghe e complesse, garantendo al contempo l'affidabilità dei risultati ottenuti.

2.4.1 Le reti per la feature extraction e per la classificazione

In questa sottosezione viene fornita una descrizione dettagliata delle architetture *Inception-v3* e *ResNet-50*, che rivestono un ruolo centrale nel prosieguo della trattazione. Queste reti sono infatti alla base sia delle *metriche di validazione* adottate, poiché utilizzate per l'estrazione delle feature necessarie al calcolo di indici come FID, KID e PRDC, sia del *processo di classificazione* impiegato per valutare in che modo le immagini generate vengano riconosciute e distinte dai modelli pre-addestrati. L'obiettivo di questa sezione è dunque richiamare in modo sintetico ma completo le principali caratteristiche architetturali delle due reti che verranno spesso richiamate in seguito.

Inception-v3

L'architettura *Inception v3*, introdotta da Szegedy et al. (2016) [54], rappresenta un'evoluzione significativa rispetto ai modelli precedenti della famiglia Inception e ai modelli più semplici caratterizzati da un elevato numero di parametri. L'obiettivo del modello è ridurre significativamente i costi computazionali e il numero di parametri, evitando così colli di bottiglia nella propagazione dell'informazione e mantenendo al contempo una forte capacità di rappresentazione.
A tal proposito, dal punto di vista architetturale Inception v3 si distingue per l'uso sistematico della *factorization* delle convoluzioni: invece di ricorrere a filtri di grandi dimensioni, costosi in termini di operazioni, vengono adottate decomposizioni in sequenze di filtri più piccoli. Ad esempio, un filtro 7×7 viene sostituito da una serie di convoluzioni 3×3, mentre convoluzioni simmetriche vengono spesso rimpiazzate da convoluzioni asimmetriche come $1 \times n$ seguite da $n \times 1$. Questa scelta riduce significativamente il numero di parametri e di moltiplicazioni, aumentando l'efficienza senza penalizzare l'espressività della rete [54].
Un altro aspetto distintivo è l'impiego di *auxiliary classifiers*, ossia rami secondari che si innestano nei livelli intermedi della rete: inizialmente pensati per migliorare il flusso del gradiente durante l'addestramento, questi classificatori si rivelano particolarmente utili come meccanismi che contribuiscono a stabilizzare il training e a ridurre l'overfitting. A ciò si affianca l'uso di tecniche di regolarizzazione come il *label smoothing*, che riduce l'eccessiva fiducia del modello nelle proprie predizioni e ne aumenta la generalizzazione [54].
La struttura complessiva della rete (Tabella 2.4) è caratterizzata da una sequenza di blocchi Inception disposti a griglie di diversa dimensione. Questo schema progressivo consente di bilanciare profondità e ampiezza della rete, mantenendo un flusso informativo ricco e gradualmente compresso fino al classificatore finale. L'uso di tecniche di riduzione della griglia attentamente progettate, come quelle illustrate nel lavoro originale, permette di evitare colli di bottiglia informativi e garantisce un'efficace transizione tra le diverse scale di rappresentazione.

Tabella 2.4: Struttura dell'architettura Inception v3 [54].

Tipo	**Dim. / Stride** or remark	**Output**
Input	image	$299 \times 299 \times 3$
Convolution	3×3 / 2	$149 \times 149 \times 32$
Convolution	3×3 / 1	$147 \times 147 \times 32$
Convolution (padded)	3×3 / 1	$147 \times 147 \times 64$
Max Pooling	3×3 / 2	$73 \times 73 \times 64$
Convolution	3×3 / 1	$71 \times 71 \times 80$
Convolution	3×3 / 2	$35 \times 35 \times 192$
Convolution	3×3 / 1	$35 \times 35 \times 288$
$3 \times$ Inception	(Fig. 5)	$17 \times 17 \times 768$
$5 \times$ Inception	(Fig. 6)	$8 \times 8 \times 1280$
$2 \times$ Inception	(Fig. 7)	$8 \times 8 \times 2048$
Average Pooling	8×8	$1 \times 1 \times 2048$
Linear (Fully Connected)	logit	$1 \times 1 \times 2048$
Softmax	classifier	1000 classi

ResNet-50

L'architettura *ResNet-50*, introdotta da He et al. (2016) [37], è una CNN appartenente alla famiglia delle *ResNets*, caratterizzata da 50 strati. Essa rappresenta una delle varianti più diffuse delle ResNet ed è ampiamente utilizzata in compiti di classificazione e riconoscimento di immagini, grazie alla capacità di essere addestrata su grandi dataset e di raggiungere prestazioni allo stato dell'arte [55].
Il contributo principale delle ResNet è l'introduzione dei *residual blocks*, già discussi in precedenza, che affrontano il problema della degradazione delle prestazioni nelle reti profonde: infatti, attraverso le *shortcut connections*, diventa possibile addestrare reti molto più profonde senza che il gradiente si annulli, garantendo così stabilità ed efficienza nell'ottimizzazione [37].
Dal punto di vista strutturale, come si può notare dalla Tabella 2.5, ResNet-50 adotta uno schema a blocchi "*bottleneck*", composto da tre convoluzioni sequenziali: una convoluzione 1×1 per ridurre la dimensionalità, una convoluzione 3×3 per l'elaborazione delle features e infine una convoluzione 1×1 per il ripristino della dimensionalità originale. Questa struttura consente di ridurre il numero di parametri e le operazioni computazionali, mantenendo al tempo stesso un'elevata capacità di rappresentazione. Le *identity shortcut connections* (o, quando necessario, proiezioni tramite convoluzioni 1×1 per adattare le dimensioni) preservano il flusso del gradiente e semplificano la propagazione delle informazioni attraverso i numerosi strati.

Rispetto a modelli precedenti come VGG-16 o GoogLeNet, ResNet-50 riesce a raggiungere una maggiore profondità con un numero contenuto di parametri, grazie alla combinazione tra residual connections e blocchi bottleneck, rendendola un compromesso ideale tra accuratezza e costi computazionali, adatta a una vasta gamma di applicazioni. Come evidenziato anche da analisi divulgative [55], i punti di forza principali di ResNet-50 possono essere riassunti in tre aspetti: (i) la possibilità di addestrare reti molto profonde, (ii) l'efficienza derivante dai blocchi bottleneck e (iii) la flessibilità che ne ha favorito l'adozione come backbone in compiti che spaziano dalla classificazione alla segmentazione e generazione di immagini.

Tabella 2.5: Struttura architetturale della ResNet-50 [37]. Ogni blocco residuale ha struttura *bottleneck* composta da convoluzioni 1×1, 3×3 e 1×1.

Stage	**Output size**	**Struttura**
Conv1	112×112	7×7, 64, stride 2
MaxPool	56×56	3×3, stride 2
Conv2_x	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Conv4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
Conv5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
Average Pool	1×1	Global Avg Pool
FC + Softmax	$1 \times 1 \times 1000$	1000 classi ImageNet

2.4.2 Analisi e calcolo delle metriche di valutazione esterne

La valutazione oggettiva della qualità delle immagini generate rappresenta un aspetto cruciale nello sviluppo e nella validazione dei modelli generativi, che può essere affrontato mediante metriche specifiche per le GAN. Queste considerano due dimensioni fondamentali: la fedeltà, ossia quanto le immagini sintetiche risultano simili a quelle reali, e la diversità, ovvero la capacità del modello di rappresentare in maniera ampia la distribuzione dei dati di partenza [53].

Nel contesto della sintesi di immagini dermoscopiche, la valutazione non può limitarsi a giudizi soggettivi, soprattutto in assenza di pareri diretti di esperti dermatologi come nel nostro caso, ma richiede indicatori quantitativi in grado di correlare qualità visiva e rilevanza clinica. Tra le metriche più impiegate si annoverano il *Fréchet Inception Distance* (FID) [49], il *Kernel Inception Distance* (KID) [50], l'*Inception Score* (IS) e le misure basate su *Precision–Recall–Density–Coverage* (PRDC) [51]. Pur basandosi su approcci differenti, tutte mirano a fornire una valutazione standardizzata e riproducibile della qualità delle immagini GAN, risultando strumenti essenziali per analisi comparative e ottimizzazione dei modelli.

Per il calcolo delle metriche, nel presente studio sono state considerate *5000 immagini reali* e *5000 immagini sintetiche*, sottolineando che il loro valore numerico finale può variare sensibilmente in funzione del numero di campioni utilizzati: un numero maggiore di immagini consente in generale una stima più stabile e affidabile. Tuttavia, disponendo di poco più di 5000 immagini reali di melanoma, questa scelta ha rappresentato un compromesso necessario, che garantisce comunque risultati statisticamente significativi e coerenti con la letteratura.

Infatti, ciascuna metrica è stata calcolata in tre configurazioni differenti:

- considerando esclusivamente melanomi reali e sintetici, per valutare la fedeltà delle immagini generate rispetto al dominio patologico;

- considerando solo nevi benigni reali e sintetici, per analizzare la qualità della generazione nel dominio sano;

- considerando un mix bilanciato di 2500 immagini reali e 2500 sintetiche per ciascun dominio, al fine di includere anche l'aspetto legato alla diversità complessiva delle immagini generate.

In quest'ultima configurazione bisogna considerare che la varietà osservabile è naturalmente limitata, poiché il modello è stato addestrato su soli due domini (nevi e melanomi), a differenza di altri studi in cui le GAN vengono impiegate per generare immagini appartenenti a molteplici classi. È stata inoltre prestata particolare attenzione a evitare che nel dataset di confronto fossero presenti coppie corrispondenti (ad esempio un nevo reale e la sua versione generata), in modo da non introdurre correlazioni spurie che avrebbero potuto alterare le stime di similarità tra i due insiemi di immagini.
In seguito, viene presentata una trattazione teorica delle metriche adottate, con l'obiettivo di illustrarne i principi e la formulazione matematica, e di descrivere in modo chiaro come esse siano state calcolate operativamente in questo studio.

Fréchet Inception Distance (FID)

Il *Fréchet Inception Distance* (FID) è una delle metriche più utilizzate per la valutazione quantitativa della qualità delle immagini generate da modelli generativi, introdotta da Heusel et al. (2017) [49]. Il principio alla base del FID consiste nel confrontare le distribuzioni delle feature estratte da immagini reali e sintetiche tramite una rete pre-addestrata, tipicamente *InceptionV3*. Le feature, rappresentate come vettori ad alta dimensionalità, vengono modellate come distribuzioni gaussiane multivariate, e il FID ne misura la distanza di Fréchet, fornendo una stima congiunta di fedeltà e diversità delle immagini generate.

Formalmente, il FID è definito come:

$$\mathrm{FID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{2.4}$$

dove μ_r e Σ_r indicano rispettivamente media e matrice di covarianza delle feature delle immagini reali, mentre μ_g e Σ_g rappresentano le corrispondenti statistiche delle immagini generate. Il termine $\mathrm{Tr}(\cdot)$ indica la traccia di una matrice. Valori di FID più bassi indicano una maggiore somiglianza tra le distribuzioni; punteggi inferiori a 15 sono generalmente associati a immagini di alta qualità, mentre valori superiori a 100 suggeriscono problemi come *mode collapse* o artefatti percettibili [49, 29].

L'*affidabilità* del FID in ambito medico dipende fortemente dallo spazio delle feature scelto: l'uso di reti addestrate su immagini naturali può introdurre bias e ridurre la sensibilità verso variazioni morfologiche clinicamente rilevanti [53]. In dermatologia, il FID è spesso utilizzato come principale metrica quantitativa per valutare la fedeltà delle immagini dermoscopiche sintetiche di melanomi e nevi. Ad esempio, nel recente lavoro di Luschi et al. (2025) [29] è stato ottenuto un valore minimo di 18.89 con StyleGAN2 per le immagini generate di melanoma, suggerendo che un FID intorno a 18 possa rappresentare un benchmark di riferimento.

Tuttavia, il FID presenta *limiti* importanti: può risultare fuorviante in presenza di *data leakage* o duplicazione interna dei campioni e non è sempre sensibile a alterazioni morfologiche localizzate, come margini tumorali o texture dermica [29, 53]. Per questo motivo, studi recenti propongono framework di validazione multifattoriali che combinano metriche quantitative, test su compiti downstream (es. segmentazione o classificazione) e valutazioni qualitative da parte di esperti clinici. Questo approccio è essenziale per garantire l'*adequacy-for-purpose* delle immagini sintetiche in contesti medici sensibili, dove la sola somiglianza statistica non è sufficiente a garantire validità anatomica o diagnostica.

Per il calcolo del FID è stata utilizzata una funzione già implementata nella libreria *PyTorch*, che richiede in input semplicemente i percorsi delle cartelle contenenti rispettivamente le immagini reali e quelle generate, e restituisce in output il valore

numerico del FID, che quantifica la distanza tra le distribuzioni delle feature estratte dai due insiemi di immagini.

Kernel Inception Distance (KID)

Il *Kernel Inception Distance* (KID) è una metrica introdotta come alternativa al FID per superarne alcune limitazioni metodologiche e statistiche [50]. Come il FID, confronta le distribuzioni delle feature estratte tramite *InceptionV3* da immagini reali e sintetiche, ma non assume che queste distribuzioni siano gaussiane, bensì si basa sul concetto di *Maximum Mean Discrepancy* (MMD), una misura non parametrica della distanza tra distribuzioni di probabilità nello spazio delle feature. Formalmente, il KID è definito come:

$$\mathrm{KID}(X_r, X_g) = \mathbb{E}[k(x_r, x_r')] + \mathbb{E}[k(x_g, x_g')] - 2\mathbb{E}[k(x_r, x_g)], \tag{2.5}$$

dove $x_r, x_r' \in X_r$ e $x_g, x_g' \in X_g$ sono i vettori di feature delle immagini reali e generate, e $k(\cdot, \cdot)$ è un kernel polinomiale:

$$k(x, y) = \left(\frac{1}{d} x^\top y + 1\right)^3, \tag{2.6}$$

con d pari alla dimensione dello spazio delle feature. Il KID misura quindi la distanza tra le medie delle distribuzioni nello spazio di riproduzione a kernel (RKHS: *Reproducing Kernel Hilbert Space*), fornendo una stima non biasata della divergenza tra distribuzioni.

Questa caratteristica lo rende più stabile e statisticamente consistente in presenza di dataset di dimensioni ridotte, condizione comune in ambito medico, rispetto al FID, che risulta sensibile alla varianza campionaria e alle assunzioni gaussiane. Valori di KID prossimi a zero indicano elevata somiglianza tra distribuzioni reali e sintetiche, mentre valori più alti denotano maggiore divergenza e qualità inferiore delle immagini generate. In dermatologia computazionale, il KID è spesso utilizzato in combinazione con il FID, integrando sensibilità alle differenze globali e maggiore robustezza statistica [28, 53]. Ad esempio, Luschi et al. (2025) riportano un $KID = 0.0025$ per StyleGAN2, indicativo di una stretta corrispondenza tra distribuzioni reali e generate.

Tuttavia, anche il KID presenta limiti: come il FID, risulta poco sensibile a variazioni morfologiche localizzate o ad alterazioni clinicamente rilevanti, e le stime possono variare significativamente a seconda della rete utilizzata per l'estrazione delle feature [53].

Per il *calcolo* del KID è stata utilizzata la funzione già implementata nella libreria *TorchMetrics*, che consente di stimare la distanza tra le distribuzioni delle feature delle immagini reali e di quelle generate. La funzione accetta come input

un estrattore di feature e nel nostro caso è stato impiegato il livello 2048 della rete Inception-v3, corrispondente alla configurazione predefinita della funzione, in quanto garantisce una rappresentazione ad alta dimensione utile per confrontare le distribuzioni di immagini complesse.
Inoltre, sono stati mantenuti i parametri di default per i restanti argomenti della funzione, tra cui:

- *subsets = 100*, ovvero il numero di sottoinsiemi su cui viene calcolata la media e la deviazione standard dei punteggi;
- *subset_size = 1000*, ovvero il numero di campioni selezionati casualmente in ciascun sottoinsieme.

La funzione restituisce in output la media e la varianza dei valori di KID ottenuti sui diversi sottoinsiemi, fornendo una stima robusta della distanza tra i due domini di immagini.

Precision–Recall–Density–Coverage (PRDC)

Le metriche *Precision–Recall–Density–Coverage* (PRDC) sono state sviluppate per valutare separatamente la fedeltà e la diversità dei modelli generativi [51, 56]. A differenza di altre metriche unidimensionali come FID o KID, che forniscono un singolo valore sintetico della distanza tra distribuzioni reali e generate, le PRDC permettono un'analisi più dettagliata, distinguendo tra la capacità del modello di produrre immagini realistiche (fedeltà) e la sua capacità di coprire la varietà del dataset reale (diversità).

Precision e *recall* per modelli generativi sono state introdotte da Sajjadi et al. (2018) [57] e successivamente perfezionate da Kynkäänniemi et al. (2019) [51].
Siano $\{X_i\}$ e $\{Y_j\}$ i campioni estratti rispettivamente dalla distribuzione reale e dal modello generativo; siano $\{f(X)\}$ e $\{f(Y)\}$ gli insiemi "embedded" di feature estratte da immagini reali e sintetiche tramite una rete pre-addestrata: su questi insiemi vengono costruiti due *manifold* nello spazio delle feature, uno per le immagini reali e uno per quelle generate, basati sulle distanze dei k-nearest neighbours (k-NN).

La *precision* misura la frazione di campioni sintetici che ricadono all'interno del manifold delle immagini reali, rappresentando la fedeltà del modello, mentre la *recall* quantifica la frazione di campioni reali coperti dal manifold generato, indicativa della diversità:

$$\text{Precision} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{Y_j \in \text{manifold}(X_1 \ldots X_N)}$$

$$\text{Recall} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{X_i \in \text{manifold}(Y_1 \ldots Y_M)}$$

dove N e M sono il numero di campioni reali e immaginari. La funzione indicatrice assume valore 1 se un campione appartiene al manifold definito dai suoi k vicini più prossimi.
I manifold sono definiti formalmente come:

$$\text{manifold}(X_1, \ldots, X_N) := \bigcup_{i=1}^{N} B(X_i, \text{NND}_k(X_i))$$

dove $B(x, r)$ è la sfera in $\mathbb{R}^D$ centrata in x con raggio r e $\text{NND}_k(X_i)$ denota la distanza dal k-esimo vicino più prossimo. Queste metriche consentono di rilevare comportamenti anomali nei modelli generativi, quali l'*overfitting* (alta precision, bassa recall) o il *mode collapse* (bassa recall).

La costruzione dei manifolds dei vicini più prossimi deve essere effettuata con attenzione, poiché le sfere attorno a ciascun campione non sono normalizzate in base ai loro raggi o alla densità relativa dei campioni nel vicinato. Per superare questo tipo di limitazione di precision e recall, Naeem et al. (2020) [56] hanno introdotto *density* e *coverage*, che migliorano la robustezza della valutazione, come viene ben illustrato in Figura 2.12:

Density estende il concetto di precisione, non solo considerando se un campione sintetico appartiene al manifold reale, ma anche quantificando quante regioni di densità locale lo contengono, risolvendo l'overestimation del manifold attorno agli outliers reali:

$$\text{Density} = \frac{1}{kM} \sum_{j=1}^{M} \sum_{i=1}^{N} \mathbf{1}_{Y_j \in B(X_i, \text{NND}_k(X_i))}$$

Coverage migliora la metrica di recall per quantificare meglio questa misura, costruendo i manifolds dei vicini più prossimi attorno ai campioni reali anziché ai campioni sintetici, poiché questi ultimi hanno meno outlier. Fornisce una misura stabile della diversità e riduce il costo computazionale, rappresentando, quindi, la frazione di campioni reali che sono coperti da almeno un campione sintetico:

$$\text{Coverage} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\exists j : Y_j \in B(X_i, \text{NND}_k(X_i))}$$

Nel dominio delle immagini dermoscopiche, le metriche PRDC risultano particolarmente efficaci nel distinguere tra modelli che generano immagini realistiche ma poco variabili (alta *precision* e *density*, bassa *coverage*) e modelli che riproducono un'ampia varietà di lesioni ma con minore fedeltà visiva (alta *recall* e *coverage*, bassa *precision*). Tuttavia, rimangono sensibili alla scelta dello spazio delle feature e al parametro k, e il loro calcolo può risultare computazionalmente oneroso per dataset di grandi dimensioni [53].

Anche per il calcolo della metrica PRDC è stato possibile utilizzare una funzione "*compute_prdc*", descritta nel lavoro di Naeem et al. (2020) [56], già implementata in Python e disponibile tramite il pacchetto "*prdc*". La funzione calcola le quattro componenti della metrica a partire dalle feature estratte dai due insiemi di immagini, reali e sintetiche, e dal parametro k che rappresenta il numero di vicini più prossimi (nearest neighbours) considerati per ciascun punto nello spazio delle feature per la definizione dei manifold. La rete utilizzata per l'estrazione delle features è la ResNet-50, vista la sua efficienza e la sua stabilità già discusse; k è stato fissato a 5, valore di default per la funzione, poiché, come mostrato da Naeem et al. (2020) [56], rappresenta un buon compromesso tra stabilità numerica e sensibilità della metrica, ed è risultato efficace in numerose applicazioni di valutazione delle GAN.

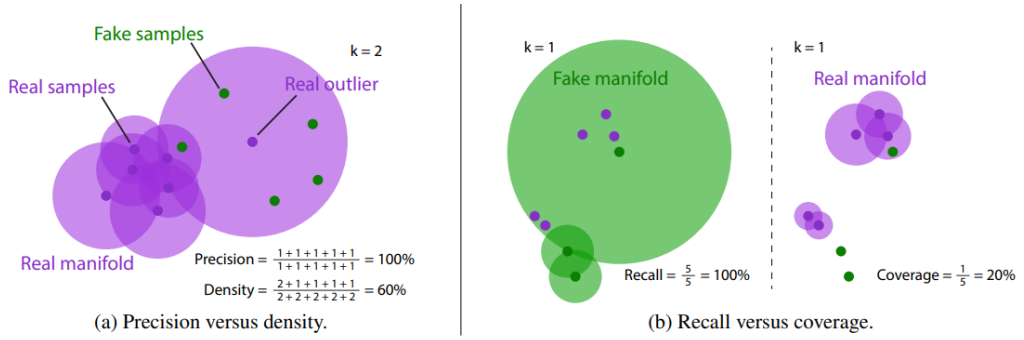


Figura 2.12: Due scenari esemplificativi che mostrano il vantaggio di utilizzare *density* rispetto a precision e *coverage* rispetto a recall nella valutazione di modelli generativi [56].

2.4.3 Sintesi comparativa e benchmark dalla letteratura

La Tabella 2.6 riassume le principali caratteristiche delle metriche quantitative adottate per la valutazione dei modelli generativi nel dominio della sintesi di immagini dermatologiche.
Come benchmark, dalla letteratura visionata riguardante la generazione di immagini di lesioni cutanee si riportano i seguenti risultati: SL-StyleGAN con *Precision* 0.525 e *Recall* 0.220 come miglior risultato in [24]; più recentemente in [29] DCGAN con *Precision* 0.7500 e *Recall* 0.8125; StyleGAN2 con *Precision* 0.7408 e *Recall* 0.2604 [29]. In particolare, Luschi et al. (2025) [29] evidenziano come StyleGAN2 ottenga valori migliori di FID (18.89) e KID (0.0025), indicando immagini più realistiche e fedeli, mentre DCGAN, con FID (203.77) e KID (0.0660) peggiori, mostra maggiore diversità (recall più elevata) ma qualità visiva inferiore.

Possiamo affermare sulla base della letteratura analizzata che la combinazione di PRDC con FID e KID rappresenta attualmente uno degli approcci più completi

per la valutazione delle immagini sintetiche, riducendo i bias e fornendo una stima più affidabile della *adequacy-for-purpose* dei modelli generativi in ambito medico.

Tabella 2.6: Confronto tra le principali metriche di valutazione.

Metrica	Principio di calcolo	Punti di forza	Limiti principali
FID	Distanza di Fréchet tra le distribuzioni gaussiane multivariate delle feature di immagini reali e sintetiche.	Valuta fedeltà e diversità simultaneamente; robusto e ampiamente accettato; buona correlazione con la percezione visiva.	Assume distribuzioni gaussiane; sensibile alla dimensione del campione e al dominio delle feature; non rileva variazioni locali.
KID	Basato sulla *Maximum Mean Discrepancy* (MMD) con kernel polinomiale di grado 3; non assume gaussianità.	Stima non biasata; più stabile del FID su dataset ridotti; adatto ad applicazioni mediche con dati limitati.	Dipende dalla scelta del kernel e della rete di feature; bassa sensibilità a variazioni morfologiche localizzate.
PRDC	Valuta la sovrapposizione e la copertura tra i manifold delle feature di immagini reali e sintetiche.	Permette di distinguere fedeltà e diversità; correlato con la percezione umana; sensibile alla distribuzione locale.	Richiede scelta di soglie (τ); dipende dallo spazio delle feature; computazionalmente più oneroso.

2.4.4 Utilizzo delle metriche con reti riaddestrate

Le reti utilizzate nel calcolo delle metriche sono pre-addestrate su ImageNet, un dataset che include milioni di immagini, ma nessuna di queste è dermoscopica. Di conseguenza, sebbene tali metriche siano comunemente riportate nella letteratura, abbiamo ritenuto che i risultati ottenuti potessero essere parziali o fuorvianti. Per questo motivo, abbiamo deciso di utilizzare un approccio innovativo basato sul transfer learning per effettuare un *riaddestramento delle reti Inception-v3 e ResNet50*, in modo tale da permettere a tali reti di estrarre le caratteristiche non solo in base a ciò che hanno appreso da ImageNet, ma anche in relazione alle specifiche peculiarità delle immagini di lesioni cutanee.
Sono state poi ridefinite e implementate le funzioni per il calcolo delle metriche usando le reti pre-addestrate.
Questo approccio ci permette di avere un miglior allineamento delle metriche con

il task specifico della valutazione delle lesioni generate, da cui ci aspettiamo di ottenere valori più coerenti e consistenti con l'obiettivo di questo lavoro.

Nel seguito sono descritti i dettagli relativi all'addestramento delle reti e ai risultati ottenuti nella classificazione delle immagini del validation set. L'obiettivo è sviluppare modelli in grado di apprendere efficacemente le caratteristiche distintive delle lesioni e di identificare correttamente i nevi benigni e i melanomi, sfruttando il pre-addestramento effettuato su ImageNet. Questo approccio, basato sul *transfer learning*, consente non solo di risparmiare tempi e risorse evitando la costruzione di un modello da zero, ma anche di migliorare significativamente le prestazioni.
Nel transfer learning, infatti, si evita di addestrare una rete da zero, cosa che richiede un dataset di grandi dimensioni, per utilizzare una rete pre-addestrata su un ampio dataset, come ImageNet, adattandola al task specifico.
Le principali modalità di transfer learning sono:

- **Fine-tuning della ConvNet**: la rete pre-addestrata viene utilizzata come punto di partenza, e il resto dell'addestramento prosegue come di consueto.
- **ConvNet come estrattore di caratteristiche fisso**: i pesi di tutti i layer vengono congelati, eccetto l'ultimo layer completamente connesso, che viene sostituito con uno nuovo e addestrato.

Per il presente lavoro è stato adottato un *approccio di fine-tuning parziale*, in cui sono stati aggiornati solo gli ultimi blocchi convoluzionali delle reti pre-addestrate su ImageNet. L'ultimo layer è stato sostituito con un classificatore binario, mentre i pesi dei primi strati sono stati mantenuti congelati, in modo da preservare le conoscenze generali acquisite su ImageNet e, al contempo, adattare la rappresentazione interna della rete alle specifiche caratteristiche delle immagini dermoscopiche.
In particolare, sono state scelte le architetture descritte nella Sezione 2.4.1:

- **Inception-v3**: scelta in quanto nella sua versione pre-addestrata su ImageNet è alla base del calcolo delle metriche FID e KID.
- **ResNet50**: selezionata per la sua architettura particolarmente efficace, che consente di ottenere ottimi risultati sia come estrattore di caratteristiche per il calcolo della PRDC, sia per la classificazione.

Il training è stato impostato seguendo il tutorial ufficiale di PyTorch sul "Transfer Learning for Computer Vision" [58]. Il modello ResNet50 è stato addestrato su un dataset di 10.000 immagini, suddivise equamente tra melanomi e nevi benigni, con un set di validazione di 2.000 immagini (1.000 per ciascuna classe). Per Inception-v3, il numero di immagini di lesioni benigne è stato aumentato a 7.500, a causa delle difficoltà iniziali osservate nel classificarle correttamente. Entrambi i modelli

sono stati addestrati con un batch size di 32 per 50 epoche.
Per migliorare la generalizzazione e aumentare la variabilità dei dati, è stata applicata la data augmentation. Le immagini di training sono state trasformate con *RandomResizedCrop* (224x224 pixel per ResNet50, 299x299 pixel per Inception-v3) e *RandomHorizontalFlip* per aumentare la robustezza contro le rotazioni orizzontali. Per il set di validazione, sono state utilizzate le trasformazioni *Resize(256)* e *CenterCrop(224)* per ResNet50 e *Resize(320)* e *CenterCrop(299)* per Inception-v3, al fine di mantenere la coerenza delle dimensioni senza alterare eccessivamente le immagini. Per entrambe le architetture, sono stati applicati ToTensor e normalizzazione delle immagini, utilizzando la media e la deviazione standard predefiniti di ImageNet per allineare le immagini alla distribuzione dei dati su cui i modelli sono stati inizialmente addestrati.
L'ottimizzazione dell'addestramento è stata effettuata utilizzando *Stochastic Gradient Descent* (SGD) con un learning rate iniziale di 0.001, che è stato ridotto ogni 7 epoche tramite un LR scheduler con un *gamma* di 0.1. Il *momentum* è stato impostato a 0.9. La funzione di perdita utilizzata è stata la *CrossEntropyLoss*, adatta per la classificazione multi-classe. Durante l'addestramento, sono state monitorate la loss e l'accuratezza su entrambi i set di training e validazione.
I pesi del modello sono stati salvati basandosi sulla migliore validation loss per ogni epoca, confrontando il valore attuale con il precedente. Per ResNet50, il miglior modello è stato salvato all'epoca 26, mentre per Inception-v3 il miglior modello è stato salvato all'epoca 22.
L'accuratezza ottenuta dai due classificatori sul *validation set* di immagini reali è pari al 67% per *Inception-v3* e al 78% per *ResNet-50*. Questi risultati sono in linea con quanto atteso, considerando la maggiore profondità ed efficienza della *ResNet-50*, nonché gli ottimi livelli di performance riportati in letteratura per questa rete nelle attività di classificazione di immagini (dermoscopiche e non).

2.4.5 Valutazione attraverso i classificatori esterni

La validazione delle immagini sintetiche prodotte dal modello generativo rappresenta un passaggio cruciale per verificare la qualità e la coerenza semantica dei campioni generati rispetto ai dati reali. Quindi, si è deciso di non limitarsi all'analisi delle metriche esterne che, pur offrendo una valutazione oggettiva del realismo e della diversità delle immagini, presentano criticità quando applicate a contesti medici e diagnostici.
Per queste ragioni, abbiamo deciso di utilizzare classificatori esterni già addestrati su dataset dermoscopici, così da ottenere una valutazione indipendente e più concreta della qualità diagnostica delle immagini generate. In particolare, sono stati impiegati *Inception-v3* e *ResNet-50*, riaddestrati secondo le modalità e le motivazioni illustrate nella sezione precedente, insieme al classificatore vincitore

Parametro	Inception-v3	ResNet50
Epoche	50 (best 22)	50 (best 26)
Batch size	32	32
Immagini training	5.000 + 7.500	5.000 + 5.000
Immagini validazione	1.000 + 1000	1.000 + 1000
Horizontal Flip	Sì	Sì
Resizing	320	256
Cropping	299x299	224x224
Ottimizzatore	SGD	SGD
Learning Rate (LR)	0.001	0.001
LR Scheduler	StepLR, 7 epoche	StepLR, 7 epoche
Momentum	0.9	0.9
Gamma	0.1	0.1
Loss function	CrossEntropyLoss	CrossEntropyLoss
Accuratezza	67%	78%

Tabella 2.7: Parametri e risultati di Inception-v3 e ResNet50

della competizione *SIIM-ISIC 2020*, scelto come riferimento per la sua elevata affidabilità nella distinzione tra lesioni benigne e melanomi [59].

L'obiettivo di questo approccio è verificare quanto le immagini generate risultino realistiche e coerenti dal punto di vista diagnostico, valutando se vengano riconosciute correttamente da modelli di classificazione già consolidati.
Nella pratica, le immagini sintetiche appartenenti ai due domini sono state sottoposte ai classificatori e le loro predizioni vengono poi confrontate con quelle ottenute sulle immagini reali. Se il classificatore assegna alle immagini generate probabilità diagnostiche coerenti con la classe di destinazione, si può concludere che il modello generativo ha imparato a riprodurre in modo credibile le caratteristiche visive e morfologiche tipiche di ciascun dominio.

Questo tipo di validazione si basa sull'idea che un classificatore ben addestrato su un ampio insieme di immagini dermoscopiche reali possa essere considerato un buon indicatore della qualità semantica delle immagini sintetiche. In altre parole, se le immagini prodotte dalla *CycleGAN* riescono a "convincere" un classificatore esterno indipendente, significa che il modello è stato in grado di catturare efficacemente le caratteristiche distintive tra lesioni benigne e melanomi.
Ci si aspetta quindi che un modello generativo ben addestrato produca immagini che, se valutate dal classificatore, mostrino una distribuzione delle probabilità simile a quella osservata per i dati reali e un'elevata accuratezza rispetto alla loro etichetta di riferimento. Tali risultati costituirebbero una prova quantitativa della capacità

del modello di generare immagini realistiche e coerenti, confermando l'efficacia dell'approccio proposto.

È importante sottolineare che l'utilizzo di classificatori esterni come strumento di validazione per modelli generativi in ambito dermoscopico rappresenta un aspetto innovativo di questo lavoro. A nostra conoscenza, infatti, nessuno studio precedente ha applicato in modo sistematico questo tipo di strategia per valutare la qualità semantica e diagnostica delle immagini sintetiche. Ciò rende questo contributo uno dei punti chiave della presente tesi, aprendo una prospettiva metodologica nuova per la validazione dei modelli generativi in contesti clinici.

Classificatore vincitore SIIM-ISIC 2020

Nel perseguire l'obiettivo di costruire un sistema di valutazione solido e affidabile, affinché tale valutazione fosse significativa, è stato necessario scegliere un classificatore che fosse in grado di estrarre e riconoscere correttamente le caratteristiche visive discriminanti dei melanomi e dei nevi benigni. Per questo motivo, invece di concentrare gli sforzi sull'addestramento di ulteriori nuovi modelli di classificazione, si è deciso di impiegare un classificatore già validato e riconosciuto nella letteratura scientifica come stato dell'arte nella diagnosi automatica di lesioni cutanee.

Dopo un'analisi approfondita delle soluzioni disponibili, la scelta è ricaduta sul *classificatore vincitore della competizione internazionale SIIM-ISIC Melanoma Classification Challenge 2020* [59]. Tale modello si basa su un *ensemble di CNNs* con architetture *EfficientNet*, *ResNeXt* e *ResNeSt*, addestrate con diverse dimensioni di input e differenti strategie di validazione incrociata. L'approccio proposto dagli autori ha raggiunto un'AUC pari a 0.960 in *cross-validation* e 0.949 sulla *private leaderboard*, risultando la soluzione con le migliori prestazioni tra oltre 3300 team partecipanti.

L'architettura segue la logica classica dei modelli di classificazione basati su *transfer learning*: si parte da reti profonde pre-addestrate su ImageNet e si sostituisce l'ultimo layer con un livello adattato al numero di classi del dataset, procedendo poi con un *fine-tuning* sul dominio specifico.
Gli autori hanno impiegato una strategia di validazione a 5-fold su un dataset combinato comprendente le edizioni 2018, 2019 e 2020 dell'ISIC, in modo da ridurre l'instabilità del punteggio AUC causata dal basso numero di campioni maligni.
A differenza di una semplice classificazione binaria, il modello utilizza una codifica diagnostica più granulare, con nove categorie (*nevus*, *melanoma*, *seborrheic keratosis*, *BCC*, ecc.), da cui viene successivamente estratta la probabilità associata alla classe *melanoma*.
In alcune varianti, il modello integra anche *metadati clinici* (età, sesso, sede anatomica, dimensione dell'immagine e numero di lesioni per paziente), elaborati da

due livelli densi e concatenati alle feature visive estratte dalle CNN, aumentando la diversità e la robustezza dell'ensemble.
Durante l'addestramento sono state applicate numerose tecniche di *data augmentation* (rotazioni, riflessioni, variazioni di luminosità e contrasto, distorsioni ottiche e *cutout*) per ridurre l'overfitting.
L'ottimizzazione è stata condotta tramite *cosine annealing* con una fase di warm-up di un'epoca e un totale di 15 epoche, utilizzando GPU NVIDIA Tesla V100 in precisione mista. L'ensemble finale combina i punteggi di 18 modelli eterogenei tramite la media delle probabilità normalizzate, garantendo così un'elevata stabilità e robustezza predittiva.

I pesi di questo classificatore sono disponibili su *Kaggle*, mentre nel relativo *repository GitHub* è presente una versione specificamente progettata per l'inferenza su immagini esterne. Questa implementazione consente di personalizzare diversi aspetti dell'esecuzione, come la selezione dei *fold*, l'uso di specifici sottoinsiemi di modelli, la riduzione delle trasformazioni in input o il salvataggio di *heatmaps* di attivazione.
Nel nostro esperimento è stato scelto di utilizzare il classificatore nella sua configurazione completa, includendo tutti i fold dei 18 modelli e mantenendo attive tutte le trasformazioni di input, in modo da ottenere predizioni quanto più precise e stabili possibile.
Sono state valutate 1000 immagini sintetiche, suddivise in 500 melanomi e 500 nevi benigni, garantendo così un buon compromesso tra affidabilità statistica e tempi computazionali.

Per la valutazione dei risultati è stata utilizzata l'*Area Under the Curve (AUC)*, metrica ampiamente adottata per misurare la capacità discriminativa di un classificatore, indipendentemente dalla soglia di decisione scelta. Essa rappresenta l'area sotto la curva ROC, che mette in relazione il tasso di veri positivi (TPR) con il tasso di falsi positivi (FPR) per tutte le possibili soglie di classificazione: un valore di AUC pari a 1 indica una classificazione perfetta, mentre un valore di 0.5 corrisponde a prestazioni casuali.
Ci si attende che lo score ottenuto dal classificatore sulle immagini sintetiche si avvicini a quello ottenuto sulle immagini reali: se il modello di classificazione riesce a riconoscere correttamente le immagini generate con un'accuratezza comparabile, significa che il nostro generatore è stato in grado di riprodurre in modo efficace le feature discriminanti che distinguono le due classi di lesioni.

2.5 Ottimizzazione dei risultati

Una delle componenti più significative di questo lavoro, e che introduce un ulteriore elemento di innovatività, riguarda la scelta dei metodi di campionamento per la selezione del set di addestramento della *CycleGAN*. Ci siamo infatti interrogati sul motivo per cui, nella maggior parte dei lavori presenti in letteratura, il set di training utilizzato per la generazione di immagini sintetiche venga selezionato in modo completamente casuale, nonostante siano stati più volte evidenziati alcuni limiti e criticità dei dataset ISIC, impiegati nella quasi totalità degli studi.
A partire da questa osservazione, ci siamo chiesti se una selezione più ragionata e mirata del training set potesse condurre a risultati significativamente migliori rispetto alla baseline casuale, e sotto quali aspetti in particolare (realismo, varietà o fedeltà diagnostica, ...). Questi interrogativi ci hanno spinto a esplorare diverse strategie di campionamento, ciascuna concepita per affrontare uno specifico problema legato alla selezione delle immagini di addestramento.

Un ulteriore aspetto che ha motivato questa analisi è la considerazione che la generazione di immagini a partire da set di addestramento differenti potesse non solo migliorare la qualità visiva dei risultati, ma anche consentire di ottenere diverse possibili simulazioni dell'evoluzione di una lesione, invece di una singola rappresentazione. Poiché la varietà interna dei due domini di interesse (nevi e melanomi) è estremamente ampia, questa diversificazione risulta particolarmente utile anche in ottica applicativa: in un contesto preventivo, fornire a un paziente più possibili evoluzioni visive di una lesione potrebbe rappresentare uno strumento di supporto più realistico e informativo.

L'analisi condotta ha quindi permesso di comprendere se e in quale misura determinate strategie di selezione del set di addestramento migliorassero la qualità della generazione. Successivamente, una fase di ottimizzazione dei parametri ha permesso di affinare ulteriormente i risultati, aumentando in particolare il numero di epoche e di immagini di training. Questo ha portato alla produzione di immagini ancora più fedeli e realistiche, consentendo di individuare la configurazione che, a nostro giudizio, rappresenta il miglior risultato generativo ottenuto nel corso dello studio.
Tale procedura è stata applicata partendo dal set di training che ha fornito i risultati più promettenti nella configurazione di base, evitando così di ripetere lunghi addestramenti per tutte le tipologie di campionamento e permettendo un confronto sistematico tra le diverse strategie sulla stessa baseline. L'ottimizzazione dei parametri, quindi, si è concentrata esclusivamente sul metodo più performante, garantendo ancora una volta un buon compromesso tra accuratezza sperimentale ed efficienza computazionale.

2.5.1 Raffinamento del dataset

In questa sezione viene presentato un approfondimento teorico sulle diverse strategie di selezione delle immagini adottate a partire dal dataset precedentemente filtrato e unificato, per poi illustrare nel dettaglio le procedure di implementazione seguite. Questa analisi consente di comprendere quale tra le strategie considerate porti alla generazione di immagini più realistiche, fedeli e coerenti con le caratteristiche diagnostiche tipiche dei due domini di interesse, fornendo così una base solida per la successiva fase di ottimizzazione del modello.

Campionamento randomico

Nella quasi totalità della letteratura analizzata relativa all'utilizzo di GAN o di CycleGAN per la generazione di immagini di lesioni cutanee, e in particolare nello studio di riferimento di Jutte et al. (2024) [1], la selezione delle immagini di training avviene in modo puramente casuale. In questo approccio, il set di addestramento viene estratto senza un'analisi preventiva delle caratteristiche del dataset, trascurando così eventuali problematiche legate alla sua struttura o alla presenza di ridondanze interne.

Deduplicazione con Cosine Similarity

La presenza di immagini duplicate o molto simili rappresenta un problema rilevante nei dataset medici, poiché, come già analizzato nel Capitolo 2 relativo al dataset, può portare a fenomeni di overfitting e a una sovrastima delle prestazioni dei modelli. Tale criticità è stata riscontrata anche nei dataset ISIC 2019 e ISIC 2020, sia per la presenza di immagini identiche all'interno dei singoli insiemi, sia a causa delle fasi di preprocessing che hanno portato alla loro unione in un unico dataset. Inoltre, nella suddivisione tra training e validation set, la presenza accidentale di immagini duplicate in entrambi i sottoinsiemi rischia di alterare la valutazione del modello. Gli stessi autori Cassidy et al. (2022) [34], da cui è stata tratta l'analisi dei dataset ISIC, hanno proposto l'impiego di metriche di similarità per individuare immagini ridondanti, avvertendo tuttavia che una rimozione eccessiva può ridurre la capacità di generalizzazione del modello.

Una delle metriche più utilizzate a questo scopo è la *cosine similarity*. Nel contesto delle immagini, ogni campione viene rappresentato come un vettore di feature (ad esempio tramite una rete neurale pre-addestrata), e la similarità tra due immagini x e y è definita come:

$$\text{cosine similarity}(x, y) = \cos\theta = \frac{x \cdot y}{\|x\|\|y\|}$$

dove $x \cdot y$ è il prodotto scalare tra i due vettori e $\|x\|$, $\|y\|$ le rispettive norme euclidee. Il risultato varia tra -1 e 1: valori prossimi a 1 indicano immagini molto simili, mentre valori vicini a 0 o negativi segnalano immagini dissimili. Il principale vantaggio di questa metrica è che non dipende dalla magnitudine dei vettori, ma esclusivamente dalla loro direzione nello spazio delle feature, permettendo così di rilevare ridondanze anche in presenza di variazioni di scala o luminosità.

Un framework basato proprio sulla cosine similarity per identificare immagini eccessivamente simili prima della fase di addestramento è stato proposto da Islam et al. (2024) [60]. Anche in questo caso, gli autori evidenziano come uno dei limiti principali dei dataset medici per l'addestramento di reti generative sia la scarsa variabilità inter-classe, dimostrando che con tale approccio è stato possibile ridurre la ridondanza interna dei dataset e migliorare la capacità delle GAN di generare immagini con maggiore variabilità e valore discriminativo. A maggior ragione, in ambito dermatologico, dove nevi e melanomi possono presentare caratteristiche visive molto simili, l'impiego della cosine similarity per la deduplicazione porta ad ottenere effetti positivi sia sulla qualità delle immagini sintetiche sia sulle prestazioni dei modelli di classificazione che ne fanno uso.

Cluster sampling e Instance selection

Il *cluster sampling* è una tecnica di campionamento che consiste nel suddividere il dataset in insiemi omogenei, detti *cluster*, dai quali vengono poi selezionati campioni rappresentativi per l'analisi o l'addestramento del modello.
Come evidenziato da Yen e Lee (2009) [61], all'interno di un dataset possono coesistere gruppi con caratteristiche statistiche e morfologiche differenti: un cluster composto prevalentemente da campioni appartenenti alla classe maggioritaria tenderà a comportarsi di conseguenza, mentre uno con una maggiore concentrazione di esempi della classe minoritaria rifletterà più da vicino le proprietà di quest'ultima. Questo approccio risulta particolarmente utile in presenza di ridondanza o squilibri di classe, poiché la suddivisione in cluster consente di preservare la variabilità interna del dataset e di ridurre i rischi associati a una selezione puramente casuale e non controllata.

Tra i metodi di clustering più diffusi vi è il *k-means* (vedi Algorithm 2), che suddivide i dati in k gruppi minimizzando la varianza intra-cluster rispetto ai centroidi. Nei metodi di sotto-campionamento basati su cluster, i campioni della classe maggioritaria vengono raggruppati in k cluster, e da ciascuno si estraggono campioni in proporzione alla loro dimensione, riducendo così la perdita di informazione strutturale e mantenendo al contempo la rappresentatività del dataset [61].
Una delle principali sfide di questo approccio riguarda la scelta del numero ottimale di cluster k: un valore arbitrario può infatti produrre cluster troppo compatti o troppo dispersivi. Per affrontare questo problema, Rahman et al. (2025) [62]

Algorithm 2 Algoritmo k-means

Require: Dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ con $x_i \in \mathbb{R}^d$, numero di cluster k
Ensure: Partizione dei dati in k cluster $\{C_1, \ldots, C_k\}$ e centroidi $\{\mu_1, \ldots, \mu_k\}$

1: Inizializza casualmente i centroidi $\mu_1, \ldots, \mu_k$ (oppure con *k-means++*)
2: **repeat**
3: **Assegnazione**: ogni $x_i \in \mathcal{X}$ viene assegnato al cluster con centroide più vicino:

$$C_j \leftarrow \{x_i : \|x_i - \mu_j\|^2 \leq \|x_i - \mu_h\|^2,\ \forall h\}.$$

4: **Aggiornamento**: per ciascun cluster C_j, il centroide viene ricalcolato come:

$$\mu_j \leftarrow \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i.$$

5: **until** assenza di cambiamenti nelle assegnazioni *or* minima variazione della funzione obiettivo $J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2$

propongono l'impiego di metriche di validazione interna, come il *Davies–Bouldin Index* (DBI) e il *Silhouette Score*, che misurano rispettivamente la compattezza e la separazione tra cluster, nonché la coesione interna rispetto alla separazione esterna. Applicando tali metriche al dataset ISIC 2019, gli autori hanno individuato come ottimale il valore $k = 6$ e hanno constatato che questo tipo di strategia consente di bilanciare rigore metodologico ed esplorazione, ottenendo un campionamento che riflette la struttura intrinseca dei dati.

Un ulteriore spunto è fornito dal lavoro di DeVries et al. (2020) [63], che pur non impiegando direttamente il cluster sampling, propone un approccio innovativo di *instance selection* per migliorare la qualità delle immagini generate dalle GAN. Gli autori osservano che non tutte le istanze contribuiscono positivamente all'apprendimento: outlier e campioni rumorosi o troppo isolati possono generare instabilità e ridurre la fedeltà visiva. Per questo motivo, suggeriscono di mantenere solo le immagini con punteggi di densità superiori a una soglia nello spazio delle feature, eliminando quelle più "sparse" e meno rappresentative della distribuzione reale.
I risultati sperimentali mostrano che questa selezione comporta una leggera riduzione della diversità delle immagini generate, dovuta alla rimozione di parte della variabilità del dataset originale. Tuttavia, la perdita è ampiamente compensata da un significativo incremento della fedeltà visiva, con immagini più realistiche e coerenti rispetto a quelle prodotte utilizzando l'intero dataset. La strategia rappresenta quindi un compromesso consapevole tra diversità e qualità, particolarmente promettente in ambito medico, dove la fedeltà visiva è essenziale per garantire affidabilità clinica.

2.5.2 Confronto tramite la validazione proposta

Dopo aver descritto le diverse strategie di campionamento adottate per la selezione del set di training, l'obiettivo di questa fase è stato quello di capire in modo chiaro e sistematico quale di esse permettesse di ottenere il miglior equilibrio tra qualità visiva, coerenza diagnostica e stabilità del modello generativo.
Per farlo, per ciascuna metodologia di campionamento, compresa quella casuale utilizzata come riferimento, sono stati generati alcuni insiemi di immagini sintetiche addestrando la *CycleGAN* con la medesima configurazione di parametri. In questo modo, è stato possibile confrontare in maniera diretta i risultati, isolando l'effetto dovuto esclusivamente alla diversa composizione del set di training.

Nello specifico, sono state definite e confrontate le seguenti modalità di raffinamento del dataset, di cui è mostrato un confronto delle features estratte in Figura 2.13:

- **Campionamento randomico**, che rappresenta la *baseline* di riferimento, in quanto è la modalità comunemente adottata nella maggior parte dei lavori presenti in letteratura sul tema.
- **Deduplicazione** delle immagini identiche o eccessivamente simili, eseguita tramite il calcolo della *cosine similarity* con soglia fissata a 0.05, valore considerato standard di riferimento in letteratura, per valutare l'impatto dell'eliminazione di immagini quasi identiche tra loro.
- **Cluster sampling** basato su *K-means*, applicato secondo due differenti strategie di selezione dei campioni:
  - **Cluster sampling stratificato**: a partire dai cluster individuati dal *K-means*, è stata effettuata al loro interno una selezione casuale delle immagini, rispettando la loro distribuzione. La suddivisione in training e validation mantiene, per ciascun cluster, la stessa proporzione presente nel dataset complessivo. Il fine è valutare come una rappresentatività completa e coerente delle diverse lesioni, sia in fase di addestramento che di valutazione del modello, possa impattare sulle prestazioni del modello
  - **Cluster sampling rappresentativo**: anche in questo caso viene eseguito un campionamento stratificato, ma con un criterio di selezione basato sulla distanza dal centroide nel dominio delle feature. Per ciascun cluster individuato dal *K-means*, vengono scelti i campioni più vicini al centroide, in quanto considerati i più rappresentativi della distribuzione del cluster, senza però considerare le categorie più "isolate". Questa strategia mira a valutare come il modello si comporta includendo nel set di training gli esempi più emblematici di ogni gruppo, mantenendo comunque la proporzionalità tra i cluster.

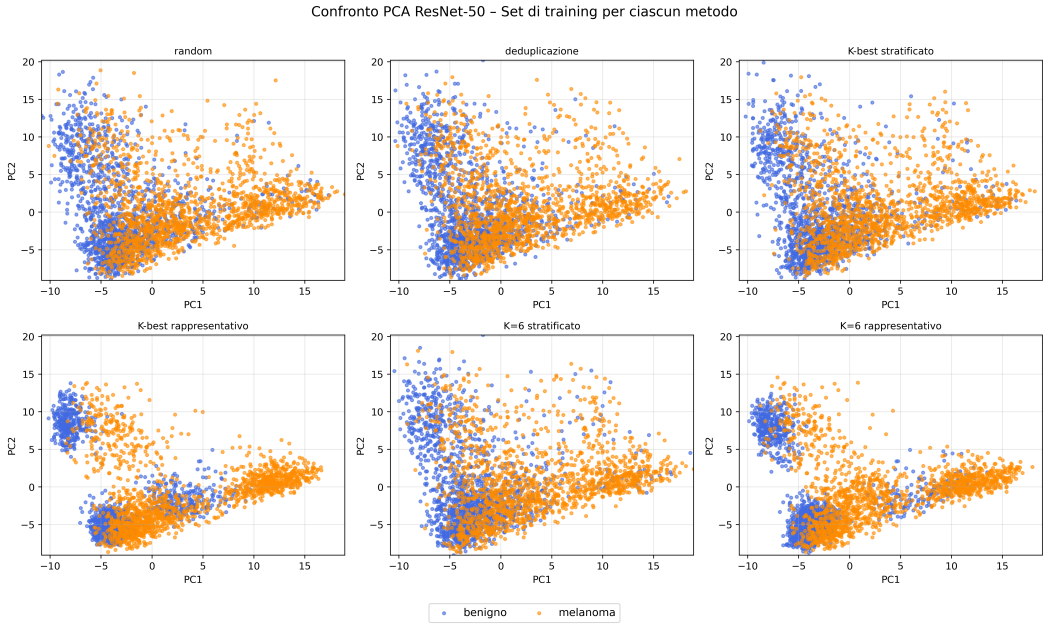
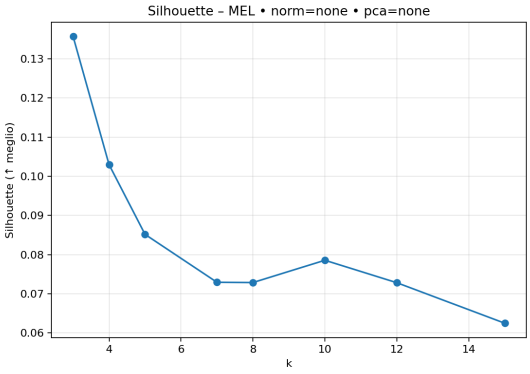


Figura 2.13: Confronto tra sei strategie di selezione del set di training visualizzate tramite PCA sulle feature estratte da ResNet-50. Ogni sottografico mostra la distribuzione delle immagini proiettate sulle prime due componenti principali (PC1 e PC2), distinguendo tra lesioni benigne (blu) e melanomi (arancione). Le strategie analizzate includono: (1) selezione casuale, (2) deduplicazione, (3) K-best stratificato, (4) K-best rappresentativo, (5) K=6 stratificato, (6) K=6 rappresentativo.

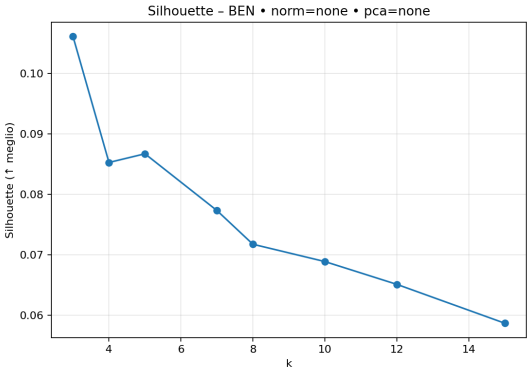
Per approfondire ulteriormente l'analisi, è stato valutato anche l'effetto della scelta del numero di cluster k sul processo di campionamento. Le due procedure di clustering descritte in precedenza sono state infatti applicate sia utilizzando il valore $k = 6$, come suggerito in letteratura da Rahman et al. (2025) [62], sia determinando k in modo ottimale sulla base di criteri interni di valutazione del clustering. In particolare, sono stati considerati il *silhouette score* e l'indice di *Davies–Bouldin*, calcolati separatamente per i due domini di interesse (melanoma e lesioni benigne).

È stato quindi deciso di individuare un valore di k specifico per ciascun dominio, in modo da eseguire il campionamento con la configurazione ottimale per entrambi i set e a tal proposito sono stati testati diversi valori di k in un intervallo compreso tra 3 e 15.
Dai grafici ottenuti (Figura 2.14), per quanto riguarda le immagini di melanoma risulta evidente che $k = 3$ rappresenta la scelta più appropriata, in quanto ottimizza simultaneamente sia il valore della silhouette sia l'indice di Davies–Bouldin.

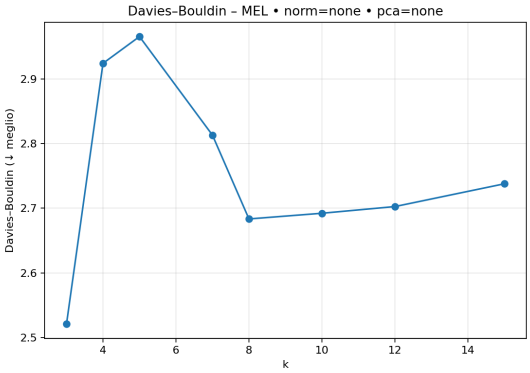
Per il dominio delle lesioni benigne, invece, la scelta risulta meno immediata: la silhouette raggiunge il massimo in corrispondenza di $k = 3$, mentre l'indice di Davies–Bouldin mostra un miglioramento netto a $k = 4$. Considerando la significativa riduzione dell'indice, la minima differenza nei valori di silhouette e l'ampiezza complessiva del dataset, è stato ritenuto più opportuno adottare $k = 4$ come valore ottimale in questo contesto.
Pertanto, nei risultati che seguiranno, quando si farà riferimento al metodo *K-best*, si intenderanno i valori di k ritenuti più adatti in base a queste analisi, ovvero $k = 3$ per il dominio dei melanomi e $k = 4$ per il dominio delle lesioni benigne. Una visualizzazione dei cluster ottenuti è riportata in Figura 2.15.



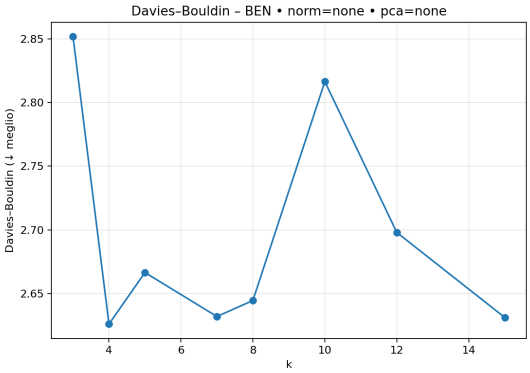
(a) Silhouette score - melanomi



(b) Silhouette score - benigni

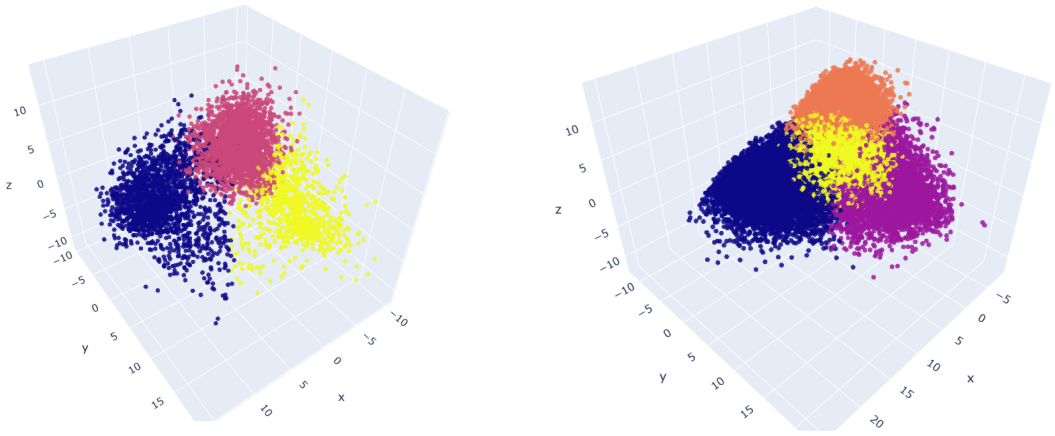


(c) Indice di Davies-Bouldin - melanomi



(d) Indice di Davies-Bouldin - benigni

Figura 2.14: Valori di *Silhouette Score* e *indice di Davies–Bouldin* per i cluster ottenuti tramite K-means, testando diversi valori di k separatamente per le immagini di lesioni benigne e melanomi. Le metriche forniscono una stima della coesione interna e della separabilità tra cluster.



(a) Melanomi - $K = 3$ **(b)** Benigni - $K = 4$

Figura 2.15: Visualizzazione tridimensionale dei cluster ottenuti tramite K-means sui due domini. I punti sono colorati in base all'appartenenza a ciascun cluster.

Una volta completato l'addestramento per ogni scenario, sono state analizzate le metriche interne del modello per valutare l'andamento delle *loss function* e la stabilità del processo di ottimizzazione. Successivamente, è stata applicata la procedura di validazione esterna descritta in precedenza, basata sia sulle metriche di similarità (FID, KID e PRDC), sia sull'impiego di classificatori esterni per verificare la coerenza diagnostica delle immagini generate.

Il confronto dei risultati ottenuti ha permesso di mettere in luce i punti di forza e le debolezze di ciascuna strategia di campionamento, evidenziando in quali casi il modello riesce a produrre immagini più realistiche o più varie e in quali, invece, tende a sovrapporsi al dominio di partenza o a perdere dettaglio diagnostico.
Nel complesso, questa analisi ha permesso di identificare quale metodologia di selezione del set di training offre le prestazioni più stabili e convincenti, fornendo un criterio utile e oggettivo per orientare la scelta del dataset nelle fasi conclusive dello sviluppo del modello. I risultati ottenuti da questo confronto sono presentati nel capitolo successivo, dedicato all'analisi e alla discussione dei risultati.

2.5.3 Estensione e ottimizzazione del training

Dopo aver completato le analisi di valutazione e il confronto tra le diverse strategie di campionamento, l'ultimo passo previsto per questa tesi è stato quello di ottimizzare ulteriormente il modello sulla base di tutte le analisi e le sperimentazioni effettuate durante questo lavoro, con l'obiettivo di generare immagini il più possibile realistiche e con valori delle metriche comparabili a quelli riportati in letteratura. Osservando in particolare l'andamento delle metriche di training e delle metriche di valutazione esterna, è emerso che il modello mostrava ancora margini di miglioramento, suggerendo la possibilità di ottenere risultati più stabili e di qualità superiore proseguendo con l'addestramento.
Per questo motivo, è stato deciso di estendere il training fino a *500 epoche*, mantenendo un *learning rate scheduler* lineare a partire dall'epoca 100, come proposto nei training del recente lavoro di Luschi et al. (2025) [29]. Inoltre, è stato incrementato il numero di immagini di training dalle 3000 della configurazione precedente a un totale di 4000 (di cui 2000 di nevi benigni e 2000 di melanomi). Tutti gli altri parametri sono stati mantenuti invariati, in quanto già associati a risultati soddisfacenti nelle fasi precedenti in relazione ai tempi di calcolo necessari. L'intento di questa fase finale è stato dunque quello di concludere il lavoro garantendo la miglior generazione di immagini possibile, sfruttando i risultati ottenuti dalle metriche interne ed esterne nella precedente configurazione e ottimizzando le risorse e i dati a disposizione.
Da questa configurazione ci si attende un miglioramento complessivo delle metriche di valutazione, sia interne che esterne. In particolare, l'ampliamento del set di training dovrebbe favorire una migliore convergenza dei valori di "M_{real}" e "M_{fake}" e una riduzione dei valori di FID e KID. Parallelamente, si prevede un incremento delle metriche di recall e coverage, grazie alla maggiore varietà e rappresentatività del dataset, e un miglioramento dei valori di precision e density, a conferma della maggiore fedeltà strutturale delle immagini generate. Inoltre, questo addestramento "finale" potrebbe tradursi in un aumento delle prestazioni dei classificatori esterni, in particolare dell'AUC ottenuta con il modello vincitore della ISIC Challenge 2020, suggerendo che le immagini sintetiche risultano non solo realistiche, ma anche clinicamente informative.

Capitolo 3

Risultati

In questo capitolo vengono presentati i risultati ottenuti a partire dalla metodologia descritta in precedenza, seguendo lo stesso percorso logico con cui si è sviluppato il lavoro, con l'obiettivo di accompagnare il lettore passo dopo passo nella scoperta dei principali esiti sperimentali.
Si apre con un riepilogo dei risultati ottenuti sul dataset, frutto delle procedure di preparazione e raffinamento adottate, insieme ai contributi specifici che questo lavoro ha introdotto in tale ambito. Successivamente, vengono analizzati i risultati dei diversi training condotti nella configurazione iniziale del modello, confrontando le varie tecniche di raffinamento del set di immagini di addestramento attraverso i sistemi di validazione implementati.
I risultati numerici delle metriche interne di addestramento e di quelle esterne di valutazione vengono riportati con grafici e tabelle, insieme a quelli ottenuti dai classificatori, e verranno giustificati sulla base delle analisi teoriche effettuate e riportate nella metodologia.
Per fornire una valutazione visiva qualitativa, vengono mostrati alcuni esempi di immagini generate a partire dalle diverse strategie di sottocampionamento e vengono commentati secondo i criteri della metrica ABCDE.
Una volta individuato l'approccio che ha prodotto le prestazioni migliori, vengono presentati i risultati ottenuti dall'estensione del training finale, includendo un confronto, per quanto possibile, con i valori di riferimento disponibili in letteratura.

3.1 Dataset

Per quanto riguarda lo studio e la costruzione del dataset di riferimento, i risultati ottenuti in questa fase riguardano principalmente l'analisi approfondita delle singole *features* dei metadati dei dataset ISIC 2019 e ISIC 2020, che è spesso assente nella letteratura, dove l'attenzione è rivolta soprattutto allo sviluppo di algoritmi addestrati con queste immagini, trascurando le caratteristiche intrinseche dei dati di partenza.
Un contributo di questo lavoro è rappresentato dallo studio dettagliato della *feature* relativa all'identificativo della lesione (*lesion ID*), oltre che dalle analisi sulle distribuzioni di età, sesso e area anatomica, anche considerate in combinazione tra loro. Queste analisi, oltre ad offrire una visione più completa del dataset e delle lesioni cutanee studiate, evidenziano l'importanza che, secondo noi, deve essere attribuita alla comprensione e alla pulizia dei dati di partenza nei lavori di questo tipo, anche per ottenere risultati complessivamente migliori.

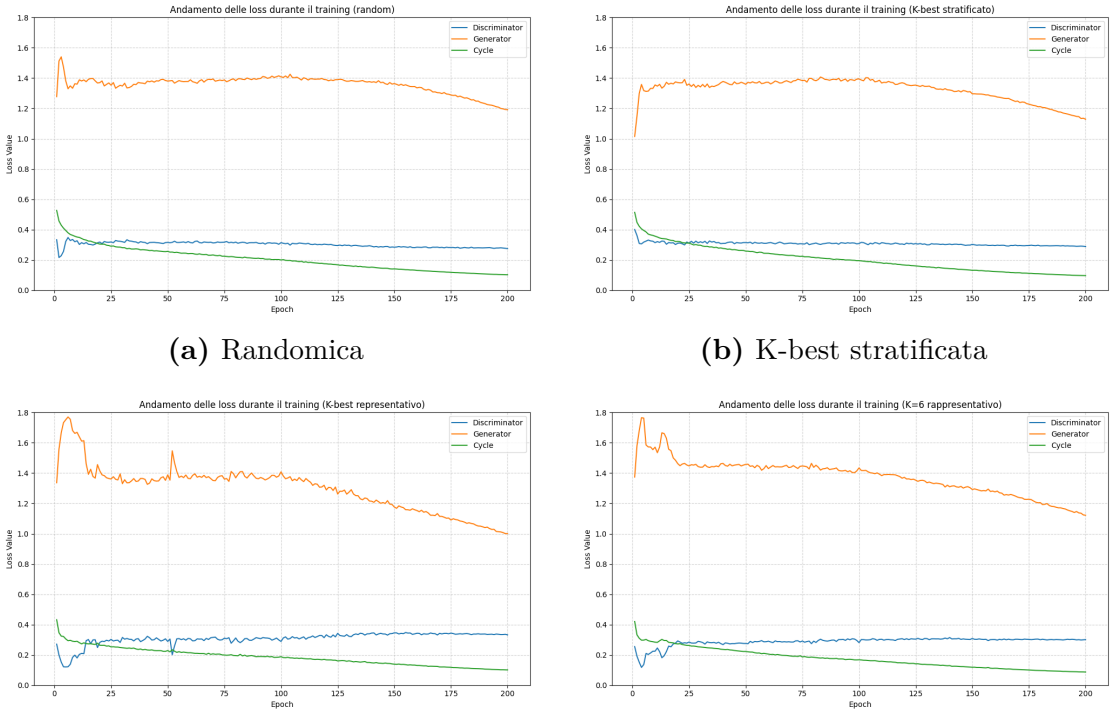
Un ulteriore risultato significativo riguarda la *creazione del dataset finale*, costruito principalmente per mitigare il problema dello squilibrio tra le classi e avere a disposizione una maggiore varietà di lesioni cutanee per le nostre analisi. L'aspetto innovativo di questo dataset risiede nel processo di filtraggio, eseguito sulla base degli studi e dei criteri diagnostici, ma soprattutto evolutivi, del melanoma, non considerati nei lavori già esistenti. Ciò conferisce al dataset una forte pertinenza rispetto al task affrontata, rendendolo una solida base per studi futuri in questo contesto.

3.2 Andamento delle metriche del training

3.2.1 Loss di training

L'andamento delle diverse componenti di *loss* nei vari metodi di campionamento analizzati mostra un comportamento complessivamente coerente con quanto riportato in letteratura sulle CycleGAN, in particolare con i risultati illustrati da [48].
Le diverse strategie di selezione del training set non modificano in modo sostanziale la dinamica di apprendimento, ma producono leggere variazioni nella stabilità delle curve e nei valori medi raggiunti.
I grafici (Figura 3.1) riportano l'evoluzione nel tempo delle tre principali componenti della *loss complessiva*: la *loss del generatore*, la *loss del discriminatore*e la *cycle-consistency loss*.

Loss del Generatore: la loss del generatore mostra un andamento decrescente, segno che il modello migliora progressivamente la propria capacità di produrre



(a) Randomica

(b) K-best stratificata

(c) K-best rappresentativa

(d) k=6 rappresentativa

Figura 3.1: Andamento delle principali funzioni di perdita (*loss*) per alcune modalità di campionamento: in arancione la *loss del generatore*, in blu quella del *discriminatore* e in verde la *cycle-consistency loss*.

immagini realistiche in grado di ingannare il discriminatore.
Nel dettaglio, all'inizio dell'addestramento si osserva in tutti i casi un picco nella loss del generatore, che raggiunge valori superiori a 1.5 nelle prime epoche, e una fase oscillatoria: ciò riflette l'instabilità della fase iniziale in cui i generatori non hanno ancora appreso pattern significativi, mentre il discriminatore riesce facilmente a distinguere le immagini sintetiche. Dopo questa fase di assestamento, la loss del generatore si stabilizza attorno a valori di circa 1.3–1.4 per un lungo intervallo di epoche (fino a circa 100–120), indicando una situazione di equilibrio provvisoria tra le due reti. Nelle ultime epoche, in concomitanza con l'avvio della linear decay del learning rate, si nota un calo progressivo della generator loss in tutti i metodi, più regolare nel campionamento stratificato, suggerendo l'efficacia del LR scheduler e un ulteriore affinamento della qualità delle immagini. Il trend discendente fino all'epoca 200 lascia inoltre ipotizzare che un proseguimento dell'addestramento avrebbe potuto portare a un ulteriore miglioramento, che non è stato apportato per rendere equo il confronto tra le diverse configurazioni e a causa delle capacità limitate di calcolo a disposizione.

Loss del Discriminatore: la loss del discriminatore rimane generalmente bassa e stabile per tutta la durata del training, indicando verosimilmente che la capacità della rete di distinguere tra immagini reali e sintetiche rimane pressoché invariata e non "domina" il generatore, che a sua volta continua a migliorare.
In particolare, dopo una fase iniziale caratterizzata da leggere oscillazioni, la curva si stabilizza su valori compresi tra 0.25 e 0.35, variando leggermente a seconda del metodo di campionamento. Questo andamento potrebbe indicare una competizione equilibrata in cui il discriminatore contribuisce a mantenere l'addestramento bilanciato.

Cycle-consistency loss La cycle-consistency loss, infine, mostra un andamento decrescente e regolare per tutta la durata del training, passando da circa 0.5 a valori prossimi a 0.1. Ciò conferma che i generatori stanno imparando a ricostruire correttamente l'immagine originale dopo un intero ciclo di traduzione ($X \to Y \to X$ e $Y \to X \to Y$).
Inoltre, una diminuzione costante di questo valore può indicare che il modello sta apprendendo trasformazioni coerenti tra i due domini, evitando mappature arbitrarie. Il fatto che la loss non si annulli completamente è positivo, poiché impedisce al modello di convergere verso una mappatura identitaria e garantisce la necessaria variabilità nelle immagini generate.

Confronto tra i metodi di raffinamento del dataset : In particolare, facendo un confronto tra i diversi tipi di raffinamento del dataset di training, osserviamo che il *campionamento casuale* (Fig. 3.1a) fornisce una baseline solida, con un andamento delle loss molto simile a quello osservato nel metodo con *deduplicazione tramite cosine similarity*.
Il *campionamento stratificato* (Fig. 3.1b) mostra invece una stabilità e una regolarità ancora maggiori in tutte le curve di loss, mentre il *cluster rappresentativo* (Figg. 3.1c, 3.1c) evidenzia una variabilità più elevata nei valori, con oscillazioni leggermente più marcate, soprattutto con "K-best", anche se riporta i valori più bassi di loss del generatore nelle fasi finali.
Una possibile spiegazione è che il campionamento stratificato garantisca la presenza di tutte le possibili varietà di lesioni presenti nel dataset, favorendo un migliore equilibrio tra apprendimento locale e capacità di generalizzazione e riducendo le fluttuazioni delle curve. Al contrario, l'addestramento su immagini di lesioni "tipiche", come nel caso del cluster rappresentativo, può portare a un apprendimento molto accurato di quei pattern specifici, ma riducendo la robustezza del modello di fronte a varianti meno comuni, generando un comportamento più instabile.

Nel complesso, la combinazione di una *generator loss stabile*, una *discriminator loss equilibrata* e una *cycle-consistency loss in progressiva diminuzione* suggerisce

che l'addestramento si sia svolto in modo *bilanciato e coerente*.
La stabilità della loss del generatore nelle prime fasi indica che il modello stia migliorando la qualità delle immagini, pur mantenendo l'equilibrio della dinamica avversaria. Nelle epoche finali, la sua riduzione segnala un ulteriore affinamento nella generazione, con immagini sintetiche sempre più simili a quelle reali e aderenti ai vincoli di coerenza ciclica.
La loss del discriminatore, invece, non segue necessariamente lo stesso andamento, poiché, nella formulazione LSGAN, è influenzata da due componenti distinte: la capacità di riconoscere immagini reali e quella di distinguere quelle sintetiche. Questo spiega perché le due curve non procedano in parallelo: il generatore ottimizza solo rispetto alle immagini generate, mentre il discriminatore bilancia il giudizio su entrambe le tipologie. Il comportamento osservato è coerente con quanto riportato nella letteratura, a conferma che entrambi i modelli mantengono una tensione avversaria efficace, senza segnali di collasso o saturazione.

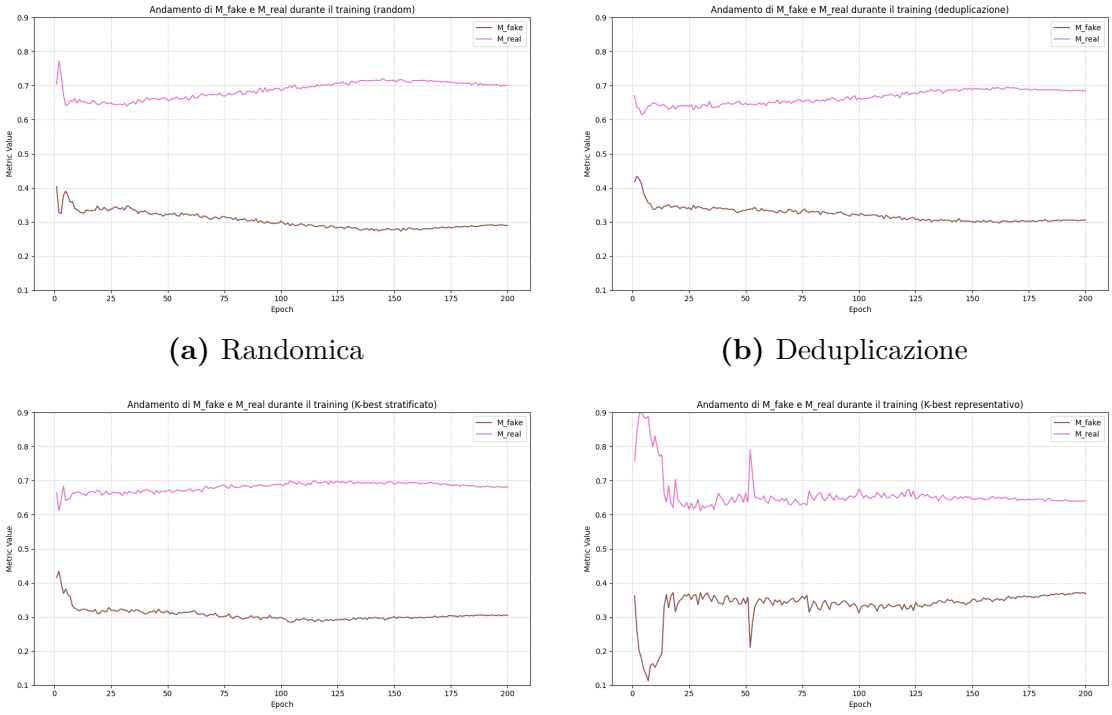
3.2.2 Output medio dei discriminatori

Analizziamo ora l'andamento dei valori "M_{real}" e "M_{fake}", osservabile in Figura 3.2, che descrivono il comportamento medio del discriminatore al variare delle epoche, rispettivamente quando riceve in input immagini reali e sintetiche del dominio melanoma.
Nel complesso, l'andamento di "M_{real}" e "M_{fake}" durante le 200 epoche di training sembra confermare in modo coerente quanto atteso: le due curve restano ben separate e mostrano una dinamica piuttosto stabile, con "M_{real}" mediamente compreso tra 0.65 e 0.70 e "M_{fake}" tra 0.28 e 0.33.

Nelle prime epoche, come previsto, si osserva una tipica fase di instabilità, caratterizzata da picchi più marcati, come osservato per la loss. Successivamente, dopo una fase in cui i due valori si distaccano leggermente o restano assestati su valori stabili, si arriva a una progressiva riduzione della distanza nelle fasi finali del training: "M_{real}" mostra una lieve decrescita, mentre "M_{fake}" decresce gradualmente, indicando che le immagini sintetiche diventano progressivamente più convincenti, ma non ancora perfettamente realistiche. Si conferma quanto già emerso dall'analisi delle loss: un prolungamento dell'addestramento potrebbe condurre a una ulteriore convergenza e quindi a un miglioramento delle prestazioni del generatore.

Passando ora al confronto tra i diversi metodi di campionamento, si osservano le seguenti differenze:

- **Campionamento randomico** (Fig. 3.2a) — L'andamento di "M_{real}" e "M_{fake}" appare regolare e coerente con una dinamica di training piuttosto stabile. Dopo le prime epoche, "M_{real}" si stabilizza attorno a 0.7 e "M_{fake}" intorno a 0.3, mantenendo una separazione costante di circa 0.4 che tende ad



(a) Randomica **(b)** Deduplicazione

(c) K-best stratificata **(d)** K-best rappresentativa

Figura 3.2: Andamento dell'output del discriminatore sulle immagini di melanoma reali ("M_{real}" - in rosa) e sintetiche ("M_{fake}" - in marrone) per alcune modalità di campionamento.

aumentare lievemente fino all'epoca 150, per poi ridursi verso la fine. Pur in assenza di strategie di selezione più sofisticate, il modello riesce a mantenere una competizione equilibrata tra le due reti, anche se la distinzione marcata tra reale e sintetico suggerisce che il generatore non abbia ancora raggiunto un livello di realismo pienamente convincente.

- **Deduplicazione tramite cosine similarity** (Fig. 3.2b) — Rispetto al campionamento randomico, "M_{fake}" resta sempre sopra 0.3 e "M_{real}" si mantiene stabilmente sotto 0.7 anche durante la fase di crescita iniziale, pertanto la distanza tra "M_{real}" e "M_{fake}" tende a essere lievemente inferiore. Questo comportamento suggerisce una potenziale miglior capacità di generalizzazione del modello, grazie alla rimozione di immagini duplicate o eccessivamente simili, che ha ridotto la ridondanza informativa del dataset, rendendo l'apprendimento più mirato.
- **Campionamento stratificato** (Fig. 3.2c) — È la configurazione che mostra la maggiore stabilità complessiva: rispetto al caso randomico, le instabilità

iniziali di "M_{real}" risultano attenuate, mentre quelle di "M_{fake}" sono leggermente più marcate ma si riducono rapidamente nelle epoche successive. Le due curve si mantengono quasi perfettamente parallele e presentano variazioni minime nel tempo, con una distanza costante inferiore a 0.4 punti, suggerendo che la suddivisione stratificata dei campioni favorisce un apprendimento più omogeneo grazie ad una rappresentazione equilibrata delle diverse categorie di immagini.

- **Campionamento rappresentativo (Fig. 3.2d)** — È la configurazione che mostra la maggiore instabilità, in particolare nelle prime 100 epoche, caratterizzate da oscillazioni ampie e irregolari, ma con la convergenza più in linea con le aspettative teoriche: con il progredire del training, la separazione tra "M_{real}" e "M_{fake}" risulta meno netta rispetto agli altri metodi, con una lenta tendenza alla convergenza verso valori prossimi a 0.5 (circa 0.64 per "M_{real}" e 0.37 per "M_{fake}"). Questo comportamento, di fatto, può avere due interpretazioni: da un lato, potrebbe riflettere un equilibrio più fragile dovuto a una maggiore incertezza del discriminatore, causata dalla limitata capacità del modello di generalizzare a immagini che si discostano dai pattern dominanti; dall'altro, invece, potrebbe indicare che il generatore produce immagini sintetiche molto simili a quelle reali, che il discriminatore fatica a distinguere. Per comprendere se la minore distanza tra i due valori derivi da un effettivo miglioramento del generatore o da un indebolimento del discriminatore, è necessario integrare questa analisi con le metriche esterne di valutazione.

Nel complesso, tutte le configurazioni hanno raggiunto una competizione sufficientemente bilanciata, ma il *campionamento stratificato* risulta essere quello più stabile e regolare, mentre la selezione *rappresentativa* introduce una variabilità maggiore e un equilibrio più delicato, sebbene con una convergenza dei valori verso 0.5 più interessante dal punto di vista teorico.
Le configurazioni con campionamento *randomico* e *deduplicazione* offrono risultati intermedi, garantendo una buona generalizzazione, ma con un controllo semantico più limitato.
Da notare infine che non sono emerse differenze significative tra le configurazioni basate su diversi valori di k nei cluster, che hanno mantenuto le stesse caratteristiche di comportamento.
Tuttavia, l'interpretazione dei soli indicatori di training rimane parziale: la natura avversaria del processo di apprendimento può infatti produrre fenomeni ambigui, motivo per cui, nelle sezioni successive, verranno analizzate le *metriche esterne di valutazione* per approfondire la qualità effettiva delle immagini generate.

3.2.3 FID nel training

Per valutare se al progredire del training corrisponda un effettivo miglioramento nella qualità delle immagini generate, oltre all'analisi delle metriche interne, è stata utilizzata anche la *FID*, la più diffusa in letteratura per misurare la somiglianza tra distribuzioni di immagini reali e sintetiche.
L'analisi, ispirata a quanto proposto da Deo et al. (2025) [53], ha previsto il monitoraggio del FID ogni 50 epoche, come mostrato nella Figura 3.3 relativa al training con selezione rappresentativa, scelta come caso esemplificativo. I risultati mostrano una diminuzione progressiva del FID, con una prima fase di calo rapido seguita da una stabilizzazione più graduale. Questo comportamento si verifica per tutti i metodi di raffinamento confrontati ed è coerente con un'evoluzione del training in cui il generatore apprende rapidamente le strutture visive principali, mentre nelle fasi successive affina la coerenza e la qualità morfologica delle immagini.

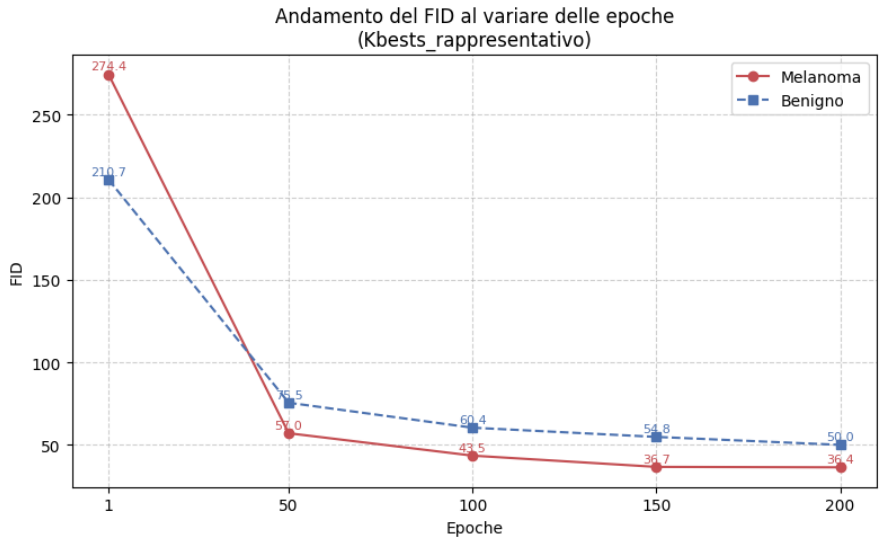


Figura 3.3: Andamento del valore di *FID* calcolato ogni 50 epoche durante l'addestramento, mostrato per il caso esemplificativo del *campionamento k-best rappresentativo*. In rosso i valori relativi alle immagini di *melanoma*, in blu quelli delle immagini *benigne*.

3.3 Metriche di valutazione esterne

In questa sezione vengono presentate le analisi relative alle metriche di valutazione esterne, utilizzate per monitorare la qualità e la fedeltà delle immagini generate e per confrontare le diverse strategie di campionamento adottate. L'obiettivo è valutare in modo oggettivo le prestazioni del modello, confrontando i risultati ottenuti con la baseline randomica rispetto a quelli derivanti dalle altre tecniche di selezione del training set e stabilire quale fra queste è la migliore.
L'analisi è articolata in due fasi principali: nella prima, sono state considerate le metriche standard calcolate mediante feature estratte da reti pre-addestrate su ImageNet, mentre nella seconda fase sono state utilizzate reti riaddestrate sul dominio dermoscopico per verificare la presenza di un eventuale miglioramento di coerenza e sensibilità delle misure rispetto alla specifica tipologia di immagini.
Ogni metrica è stata calcolata separatamente per ciascun dominio, per consentire di osservare in modo distinto come ciascun metodo di campionamento influenzi la qualità generativa e di comprendere se le differenze osservate derivino da caratteristiche intrinseche delle immagini o dalla composizione del dataset di addestramento.
In seguito, per FID e KID, i cui valori sono direttamente influenzati dalla varietà generativa, sono state effettuate analisi su un set misto, contenente immagini di entrambe le categorie, per valutare il comportamento del modello su un insieme più eterogeneo e complessivo.

3.3.1 FID

Osservando i valori ottenuti nella configurazione iniziale con 200 epoche di addestramento, i risultati, riportati in Figura 3.4 mostrano che il *Fréchet Inception Distance* (FID) oscilla complessivamente tra circa 26 e 57. In termini assoluti, questi valori risultano più elevati rispetto al *benchmark*, che indica un FID di 18.89 per immagini di melanoma [29]. Tuttavia, occorre considerare che nel lavoro citato le reti sono state addestrate per 500 epoche e con un'architettura differente e più moderna della CycleGAN, rendendo difficile un confronto diretto con i risultati in termini numerici. È ragionevole ritenere, basandoci su quanto riportato in letteratura, che i valori ottenuti in questo studio rientrino in un intervallo considerato accettabile, considerando anche la complessità del task.
Il confronto dei valori FID ottenuti per le diverse strategie di sottocampionamento evidenzia differenze significative nella qualità delle immagini sintetiche generate. In particolare, si osserva una riduzione generale del FID rispetto alla baseline casuale nella maggior parte delle configurazioni raffinate, segno che esse hanno effettivamente contribuito a migliorare la fedeltà e la varietà delle immagini generate.

Guardando invece le differenze rispetto ai domini, i valori FID per il melanoma risultano mediamente più alti rispetto a quelli del dominio benigno.

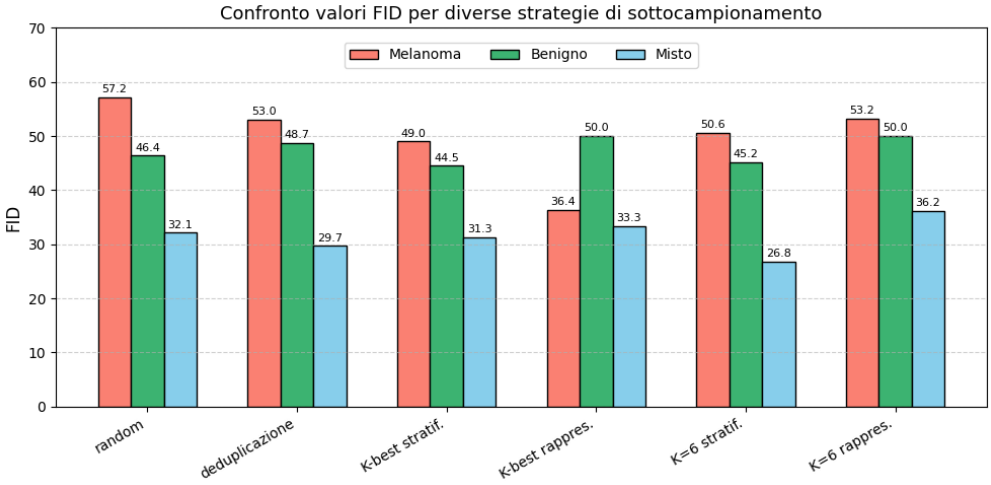


Figura 3.4: Valori di *FID* "standard" calcolati per i domini melanoma (in rosso), benigno (in azzurro) e misto (in verde).

Questo risultato può suggerire, secondo il nostro punto di vista, alcune considerazioni e due principali ipotesi:

- Il FID non misura soltanto il realismo delle immagini, ma anche la varietà statistica delle distribuzioni; pertanto, un modello che produce immagini troppo simili tra loro, anche se realistiche, ottiene comunque un valore FID elevato a causa del contributo penalizzante del termine di covarianza. Questo comportamento è coerente con la natura del dominio melanomico, caratterizzato da una maggiore omogeneità visiva rispetto alle lesioni benigne, che includono una gamma più ampia e diversificata di tipologie cutanee. Di conseguenza, il generatore apprende una distribuzione più compatta, producendo immagini meno disperse nello spazio delle feature. Questa differenza di eterogeneità è confermata anche dalla composizione dei dataset reali: circa 5.000 immagini di melanoma contro oltre 45.000 di lesioni benigne.

- Le immagini di melanoma mostrano *pattern irregolari e complessi* (asimmetrie, bordi frastagliati, variazioni cromatiche), che rendono più difficile per il generatore riprodurre fedelmente i dettagli diagnostici più rilevanti con una penalizzazione in termini di fedeltà.

Dominio melanoma. Il valore più alto si osserva per la configurazione *random* ($FID \simeq 57.2$), attribuibile all'assenza di una selezione mirata; la *deduplicazione tramite cosine similarity* riduce il valore a circa 53.0, segnalando un miglioramento moderato dovuto all'eliminazione di campioni ridondanti e alla maggiore pulizia

del dataset.
Le configurazioni basate su *cluster sampling* risultano invece le più efficaci: in particolare, il metodo *K-best representativo* ottiene il FID nettamente più basso (36.4), risultando anche l'unico caso in cui il FID del melanoma è inferiore a quello del dominio benigno. Segue la configurazione *K-best stratificata* ($FID \simeq 49.0$), che mostra comunque un miglioramento significativo rispetto alla baseline, probabilmente grazie a una maggiore diversità dei campioni inclusi nel training, che ha favorito la stabilità dell'addestramento e la generazione di immagini più varie e coerenti.

Questo risultato può essere interpretato considerando che la selezione dei campioni più rappresentativi all'interno di ciascun cluster ha favorito l'apprendimento delle *caratteristiche centrali e diagnostiche* del melanoma: il generatore, concentrandosi sulle immagini prototipiche, è probabilmente riuscito a modellare in modo più coerente la struttura semantica del dominio (forma, pigmentazione, pattern di crescita) e ad apprendere in modo più efficiente le caratteristiche tipiche del melanoma, riducendo così la distanza tra le distribuzioni reali e sintetiche. Tuttavia, un'eccessiva omogeneità dei campioni selezionati può ridurre la varietà del dataset, limitando la capacità del modello di generalizzare ai sottotipi meno comuni di melanoma.

Infine, è importante notare che anche la *scelta del numero di cluster* k sembra aver avuto un ruolo determinante nella qualità del campionamento: i risultati migliori si ottengono infatti proprio nelle configurazioni in cui k è stato determinato in modo ottimale sulla base dei criteri di validità interna del clustering, confermando la rilevanza di questa scelta nel miglioramento della generazione.

Dominio benigno. Nel dominio delle lesioni benigne si osservano valori di FID complessivamente più bassi, compresi tra circa 44 e 50. Il risultato migliore è ottenuto con la configurazione *K-best stratificata* ($FID \simeq 44.5$), seguita dalle versioni *K=6 stratificata* e random. Questa tendenza indica che, in un dominio ampio e variegato come quello dei nevi, una selezione equilibrata dei campioni nel training consente al generatore di apprendere una rappresentazione più completa della distribuzione reale, senza compromettere il realismo visivo. Le lesioni benigne presentano, infatti, caratteristiche più generiche e meno vincolanti rispetto ai melanomi: il modello ha quindi una maggiore libertà di variazione, potendo generare immagini diverse tra loro senza rischiare di uscire dai confini semantici del dominio. Questa flessibilità, combinata con la maggiore diversità rappresentativa delle immagini sintetiche indotta da tali campionamenti, riduce complessivamente il FID.
Al contrario, la configurazione *K-best rappresentativa* risulta la meno efficace ($FID \simeq 50.0$), probabilmente a causa di una selezione eccessivamente ristretta dei

campioni più centrali, che limita la capacità del generatore di riprodurre l'ampia gamma di caratteristiche visive che contraddistinguono le lesioni benigne, portando a valori FID più elevati.

Dominio misto. Il dominio "misto", che combina immagini di melanoma e di lesioni benigne, rappresenta una misura complessiva della qualità della generazione. I valori FID risultano più bassi rispetto ai singoli domini, come previsto, poiché la maggiore diversità tra le immagini reali e sintetiche contribuisce ad ampliare la distribuzione statistica di riferimento. Il miglior risultato si osserva per la configurazione *K=6 stratificata* ($FID \simeq 26.8$), segue la *deduplicazione* ($FID \simeq 29.7$) e *K-best stratificata* ($FID \simeq 31.3$); la baseline random ($FID \simeq 32.1$) mantiene comunque prestazioni discrete. Questi risultati possono essere interpretati come una via di mezzo tra i due casi precedenti: da un lato, la strategia stratificata mantiene la varietà dei campioni garantendo una rappresentatività coerente, come nel dominio benigno; dall'altro, beneficia della struttura più compatta tipica del melanoma, in cui la selezione più mirata aveva portato a un miglioramento netto.

3.3.2 KID

L'analisi dei valori di **KID** per le diverse strategie di sottocampionamento, riportati nella Tabella 3.1, conferma in gran parte le tendenze osservate con il FID, pur con alcune differenze nella scala e nella sensibilità ai metodi di selezione.
Rispetto al FID, il KID presenta alcuni vantaggi concettuali: non assume la gaussianità delle distribuzioni e fornisce, insieme al valore medio, anche una stima della sua incertezza, offrendo quindi una misura più robusta e statisticamente interpretabile.

Tabella 3.1: Valori di *KID* "standard" calcolati per i domini melanoma, benigno e misto. I valori migliori per ciascun dominio sono evidenziati in grassetto.

Metodo	**Melanoma**	**Benigno**	**Misto**
Random	0.0358 ± 0.0023	**0.0193 ± 0.0015**	0.0123 ± 0.0009
Dedupl.	0.0320 ± 0.0017	0.0228 ± 0.0012	0.0118 ± 0.0009
K-bests strat.	0.0277 ± 0.0021	0.0217 ± 0.0013	0.0127 ± 0.0010
K-bests rappr.	**0.0146 ± 0.0008**	0.0270 ± 0.0012	0.0123 ± 0.0009
K=6 strat.	0.0308 ± 0.0023	0.0227 ± 0.0015	**0.0109 ± 0.0009**
K=6 rappr.	0.0301 ± 0.0021	0.0215 ± 0.0011	0.0138 ± 0.0011

Nel complesso, i valori ottenuti si collocano in un intervallo ristretto (tra 0.010 e 0.036), indicativo di una buona similarità statistica tra immagini reali e sintetiche, ma ancora distante dal benchmark di 0.0025 [29]. Analogamente al FID, il dominio

del melanoma presenta valori mediamente più elevati, seguito dal dominio benigno, mentre il dominio misto mostra i valori più bassi e stabili; le deviazioni standard molto contenute suggeriscono una stima consistente e ripetibile.

Dominio melanoma. I valori di KID per il dominio melanomico variano tra 0.0146 e 0.0358, confermando la maggiore difficoltà del modello nel generare immagini statisticamente coerenti in un dominio con minore varietà e maggiore complessità visiva. La configurazione *random* ($KID \simeq 0.0358$) registra il valore più alto, la *deduplicazione* ($KID \simeq 0.0320$) porta a un miglioramento moderato, mentre le configurazioni stratificate si collocano in una fascia intermedia (0.0277–0.0308), garantendo prestazioni più stabili ma senza miglioramenti drastici, analogamente a quanto osservato e discusso per il FID. Il risultato più significativo si ottiene con la configurazione *K-best rappresentativa*, la quale registra il valore KID più basso dell'intero dominio (0.0146 ± 0.0008). Questo conferma l'efficacia della *selezione dei campioni più rappresentativi* nel caso di domini poco vari, che riduce la varianza interna e contribuisce a riprodurre con maggiore fedeltà le caratteristiche distintive del melanoma.

Dominio benigno. Nel dominio benigno, i valori di KID risultano complessivamente più bassi, compresi tra 0.019 e 0.027. A differenza del melanoma, la configurazione *random* ottiene qui risultati tra i migliori, mentre le configurazioni K-best stratificata e K=6 rappresentativa mostrano prestazioni analoghe ($KID \simeq 0.021\breve{}0.027$); il metodo K-best rappresentativa peggiora leggermente i risultati ($KID \simeq 0.0270$), sempre in analogia con l'andamento del FID. Questo comportamento sembra confermare ancora una volta che, in domini molto ampi e complessi, una selezione restrittiva dei campioni riduce eccessivamente la capacità del modello di catturare l'intera varietà di forme, texture e pattern cromatici.

Dominio misto. Nel dominio misto si osservano i valori KID più bassi in assoluto (0.0109–0.0138), coerentemente con l'aumento della diversità delle immagini. Il miglior risultato si ottiene con la configurazione *K=6 stratificata* ($KID \simeq 0.0109$), seguita da *deduplicazione (0.05)* e *random* ($KID \simeq 0.0123$), mentre *K=6 rappresentativa* risulta leggermente peggiore ($KID \simeq 0.0138$). Le differenze tra le medie, tuttavia, rientrano nei limiti dell'incertezza statistica, rendendo le prestazioni sostanzialmente comparabili.

Nel complesso, i risultati del KID confermano e rafforzano le conclusioni tratte dall'analisi del FID: tra i metodi, i campionamenti stratificati si dimostrano leggermente più robusti e bilanciati, soprattutto per i benigni, mentre la selezione rappresentativa produce vantaggi particolarmente rilevanti per il melanoma, dominio più limitato ma complesso.

3.3.3 PRDC e distribuzione nel feature space

Le metriche *PRDC* offrono una valutazione complementare rispetto a FID e KID, distinguendo tra la qualità visiva delle immagini (precision e density) e la loro varietà e copertura rispetto alla distribuzione reale (recall e coverage). A differenza di FID e KID, che restituiscono un singolo valore aggregato, le PRDC consentono di scomporre la valutazione in componenti specifiche, rendendo più leggibili le dinamiche del generatore. I risultati, mostrati nella Tabella 3.2 ottenuti confermano l'utilità di un approccio *multi-metrico*, capace di cogliere sfumature che sarebbero altrimenti invisibili con una sola metrica: l'analisi congiunta ha infatti restituito una visione più articolata e coerente del comportamento generativo.
Nel complesso, i valori ottenuti per queste metriche sono superiori rispetto a quelli riportati per la StyleGAN2 addestrata da [29] scelti come benchmark (*Precision* 0.7408 e *Recall* 0.2604), indicando una qualità complessiva delle immagini sintetiche e una distribuzione generata che riproduce in modo piuttosto realistico quella reale.

Metodo	**Melanoma**				**Benigno**			
	Prec.	Rec.	Dens.	Cov.	Prec.	Rec.	Dens.	Cov.
random	0.7354	0.5528	0.6384	0.5382	0.7266	0.6010	0.5647	0.5460
deduplic.	**0.7992**	0.5758	**0.7880**	**0.5896**	0.7042	0.5688	0.5059	0.5236
K-best strat.	0.7566	0.5806	0.6806	0.5846	0.7004	0.5566	0.5189	**0.5580**
K-best rappr.	0.6772	0.4006	0.5120	0.5060	**0.7638**	0.3868	**0.7998**	0.5054
K=6 strat.	0.7852	**0.5996**	0.7511	0.5882	0.7498	**0.6206**	0.5752	0.5538
K=6 rappr.	0.7266	0.5012	0.5938	0.5512	0.7046	0.3846	0.6582	0.5294

Tabella 3.2: Valori di *PRDC* "standard" per i domini melanoma, benigno e misto. I valori migliori per ciascun dominio sono evidenziati in grassetto.

Per quanto riguarda le differenze tra i due domini, le metriche di *precision* e *density* mostrano valori mediamente più alti nel dominio dei melanomi, con una differenza più marcata per la density. Questo comportamento può essere spiegato dalla maggiore compattezza e coerenza semantica del dominio melanomico: le immagini di melanoma tendono a condividere pattern visivi più simili e specifici, legati a caratteristiche cliniche ben definite. Al contrario, il dominio delle lesioni benigne è più eterogeneo, comprendendo una varietà molto ampia di sottotipi cutanei con morfologie e texture differenti. Di conseguenza, è più difficile per il generatore mantenere un'adesione stretta alla distribuzione reale, e le metriche che premiano la coerenza strutturale, come precision e density, risultano penalizzate. La *recall* appare nel complesso simile tra i due domini, mantenendosi su valori inferiori rispetto a precision e density in linea con il benchmark. La *coverage* risulta invece più bassa nei benigni (0.35–0.36) rispetto ai melanomi (0.50–0.60), suggerendo che

nel primo caso le immagini sintetiche coprono meno efficacemente l'intera variabilità reale, verosimilmente a causa della maggiore dispersione dello spazio delle feature.

Confrontando le diverse strategie di sottocampionamento emerge, seppur non in modo sistematico, un leggero *trade-off* tra *precision* e *recall*: un aumento della precision tende a ridurre la copertura e viceversa. Questo comportamento, tipico dei modelli GAN e già documentato in letteratura [63], riflette la tensione tra *fedeltà e varietà*: un generatore molto accurato tende a concentrarsi su campioni centrali, migliorando la qualità visiva ma riducendo la diversità; al contrario, una maggiore dispersione migliora la copertura, ma introduce campioni meno realistici.

Nel complesso, i risultati degli effetti di ciascuna strategia di campionamento mostrano che:

- la strategia *random* rappresenta una baseline solida ma non ottimale, con prestazioni equilibrate su tutte le metriche;
- la *deduplicazione* introduce miglioramenti modesti e riporta complessivamente i valori più alti sul melanoma grazie alla rimozione della ridondanza, ma senza incidere particolarmente sul dominio delle lesioni benigne;
- le configurazioni *stratificate* offrono nel complesso un buon compromesso, mantenendo valori di *precision*, *recall*, *density* e *coverage* bilanciati, grazie alla rappresentatività proporzionale dei campioni;
- le configurazioni *rappresentative* evidenziano un comportamento contrastante nei due domini, mostrando vantaggi significativi nei melanomi ma risultati meno efficaci nei benigni, come approfondito in seguito.

Dominio Melanoma

Osservando il comportamento delle metriche PRDC per le immagini sintetiche di melanoma (Figura 3.5), vediamo che la *baseline random* mostra valori equilibrati (precision $\simeq 0.73$, recall $\simeq 0.55$), mentre le prestazioni migliori si ottengono con i metodi di *deduplicazione* e *K=6 stratificato*, che raggiungono precision e density più elevate (0.79) e recall e coverage intorno a 0.59. La deduplicazione offre un leggero vantaggio in precision, mentre lo stratificato mostra risultati superiori in recall, come atteso, grazie alla maggiore diversità interna del campione.
Un risultato degno di nota è quello del campionamento *K-best rappresentativo*, che mostra una precision inferiore rispetto agli altri metodi, in apparente contrasto con il dominio benigno e con le metriche *FID* e *KID*, dove aveva ottenuto i valori migliori. Questa discrepanza è spiegabile considerando la differenza tra metriche globali e locali: il *FID* valuta la distanza statistica complessiva tra distribuzioni, mentre le metriche *PRDC* misurano la coerenza locale delle feature, con la precision

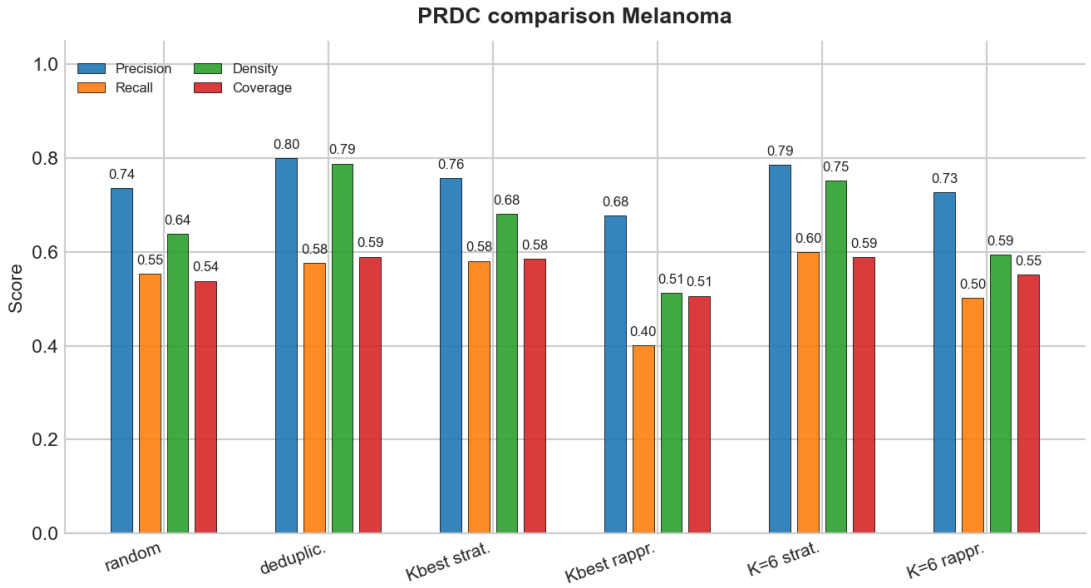
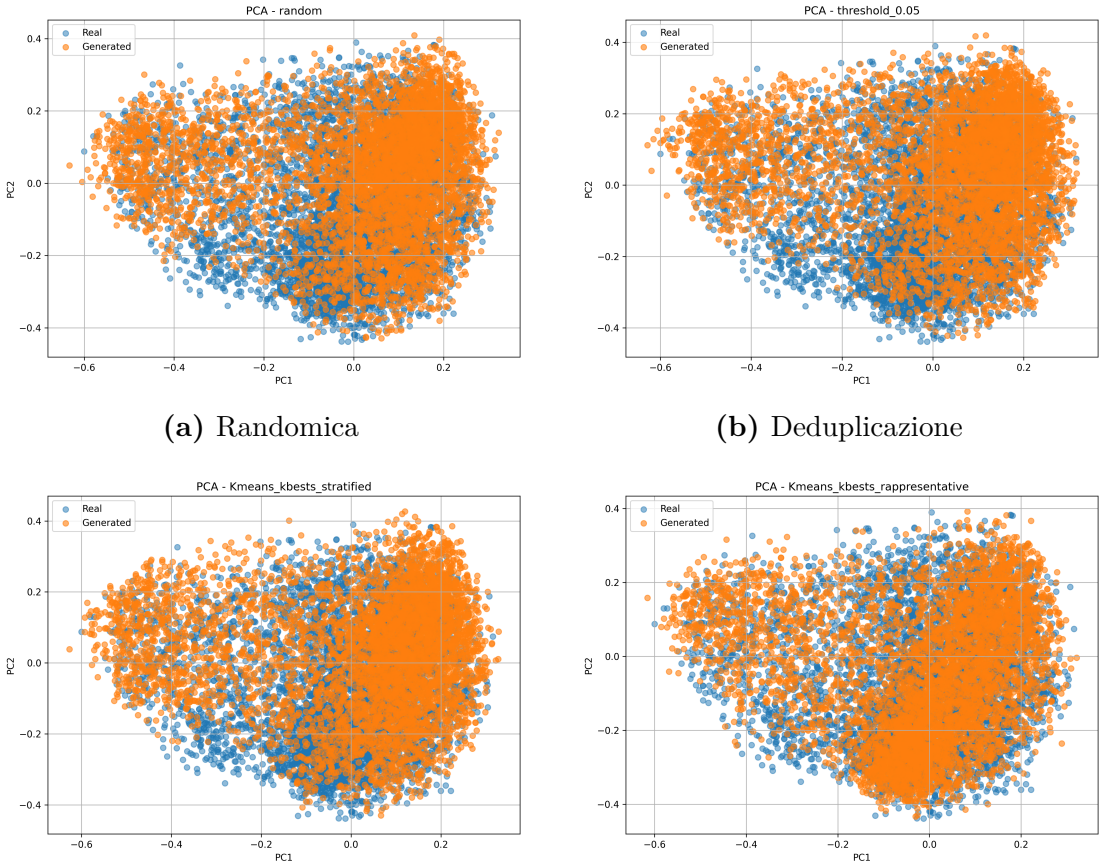


Figura 3.5: Valori delle metriche *PRDC* calcolati per il dominio *melanoma* con configurazione "standard". Le barre mostrano, per ciascun metodo di campionamento, i valori di precision (blu), recall (arancione), density (verde) e coverage (rosso).

che indica quante immagini sintetiche ricadono in regioni dense di campioni reali. Nel caso del melanoma, la selezione dei campioni più centrali ha probabilmente portato il generatore ad apprendere in modo molto concentrato le caratteristiche dominanti del dominio, producendo immagini coerenti e realistiche (da cui il FID basso), ma con una copertura limitata dello spazio delle feature (da cui *precision* e *recall* più basse). Questa dinamica riflette come la bassa variabilità intrinseca del dominio melanomico, già compatto per natura, venga ulteriormente accentuata dalla selezione rappresentativa, riducendo la capacità del modello di generalizzare ai sottotipi più rari. Si tratta di uno spunto interpretativo prezioso, difficilmente rilevabile dalla sola analisi delle metriche globali come FID e KID.
Un ulteriore fattore può essere la presenza di artefatti visivi ricorrenti, come bordature nere o segni di misurazione, che il generatore potrebbe aver enfatizzato, concentrando la distribuzione su pattern ricorrenti ma non sempre clinicamente rappresentativi.

Questi risultati trovano ulteriore conferma nei *plot delle prime due componenti principali* calcolate tramite PCA sulle feature delle immagini reali e generate di melanoma per ciascun metodo di campionamento adottato (Figura 3.6). Osservando la distribuzione delle immagini reali, si nota come le feature risultino distribuite in modo piuttosto omogeneo, con una zona di maggiore densità e senza la presenza



(a) Randomica

(b) Deduplicazione

(c) K-best stratificata

(d) K-best rappresentativa

Figura 3.6: Distribuzione delle feature delle immagini di *melanoma* reali (blu) e generate (arancione) proiettate sulle prime due componenti principali (PCA) per diverse strategie di campionamento.

di gruppi o aggregati separati. Analizzando invece i plot relativi alle immagini sintetiche, si osserva che non emergono differenze sostanziali tra i diversi metodi di campionamento, a conferma dei valori numerici complessivamente simili delle metriche PRDC.
Tuttavia, alcune tendenze possono essere notate: il campionamento randomico garantisce una discreta copertura del dominio, ma le features appaiono maggiormente concentrate nella zona di massima densità, mostrando una leggera perdita di varietà rispetto alle immagini reali. Questo comportamento si osserva in maniera più marcata anche nel campionamento rappresentativo, in cui le feature generate tendono comunque ad addensarsi maggiormente attorno al nucleo centrale della distribuzione, coprendo comunque in parte il resto del dominio.

Al contrario, i metodi deduplicazione e cluster sampling stratificato mostrano una distribuzione più equilibrata, in cui la zona di maggiore densità è popolata in modo più diffuso a favore delle regioni dello spazio circostanti.

Dominio Benigno

Nel dominio benigno (Figura 3.7), la configurazione random offre un buon equilibrio, con valori medi di precision e recall pari a 0.73 e 0.60. La deduplicazione comporta un lieve calo, mentre i metodi K-best rappresentativo e K=6 stratificato mostrano un incremento della precision, con differenze più marcate nella metrica di density.

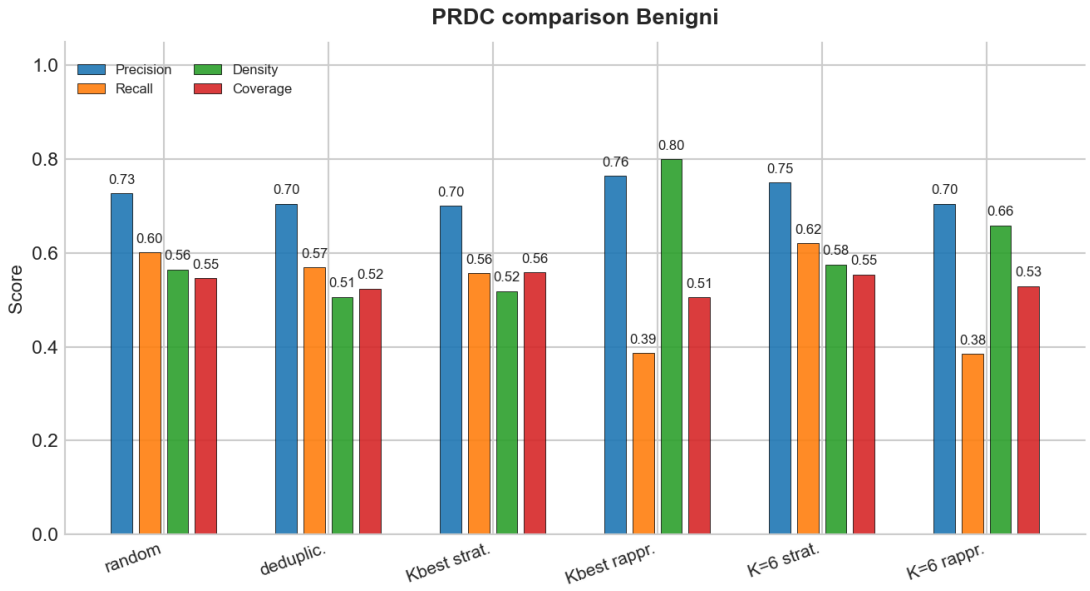
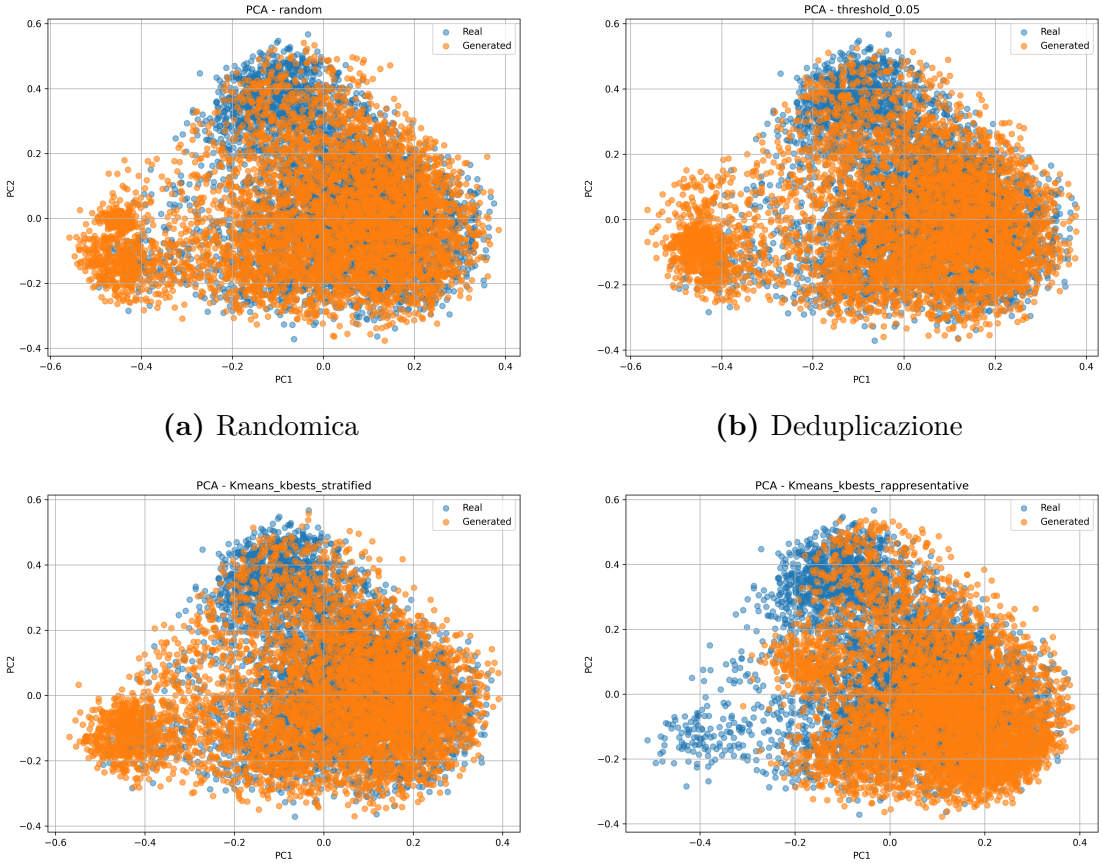


Figura 3.7: Valori delle metriche *PRDC* calcolati per il dominio dei nevi *benigni* con configurazione "standard". Le barre mostrano, per ciascun metodo di campionamento, i valori di precision (blu), recall (arancione), density (verde) e coverage (rosso).

In particolare, il *K-best rappresentativo* raggiunge il valore di density nettamente più alto (0.80), indicando che le immagini sintetiche non solo appartengono al manifold reale, ma si concentrano in regioni ad alta densità locale, dove le feature delle immagini reali sono più compatte e ricorrenti. Questo suggerisce che il generatore ha appreso con precisione le modalità visive più frequenti e clinicamente coerenti, riproducendo le features caratteristiche con maggiore realismo e riducendo il rischio di generare immagini implausibili.

Il fatto che il metodo K-best rappresentativo sia stato penalizzato dal FID, nonostante l'elevata density, indica che il punteggio FID non riflette una scarsa qualità

visiva, ma piuttosto una ridotta varietà generativa. In un dominio ampio e diversificato come quello benigno, questa limitazione ha un impatto maggiore rispetto al melanoma, dove la bassa variabilità intrinseca rende il FID meno sensibile alla dispersione. Tuttavia, la density elevata può anche indicare una tendenza ad una forte concentrazione, con il modello focalizzato sulle aree centrali della distribuzione e meno capace di rappresentare le varianti periferiche. Le metriche di recall e coverage confermano in parte questa dinamica: si riducono significativamente nei metodi rappresentativi, mentre la strategia K=6 stratificata mantiene una copertura più ampia e valori più bilanciati.



(a) Randomica **(b)** Deduplicazione

(c) K-best stratificata **(d)** K-best rappresentativa

Figura 3.8: Distribuzione delle feature delle immagini di nevi *benigni* reali (blu) e generate (arancione) proiettate sulle prime due componenti principali (*PCA*) per diverse strategie di campionamento.

Anche per il dominio delle lesioni benigne, i risultati trovano conferma nei *plot delle prime due componenti principali* calcolate sulle features delle immagini reali

e generate per ciascun metodo di campionamento (Figura 3.8). In questo caso, a differenza di quanto osservato per i melanomi, emergono differenze più marcate tra le diverse strategie, in particolare tra il campionamento rappresentativo e gli altri tre metodi.
Mentre i metodi randomico, deduplicato e stratificato mostrano una copertura del dominio reale piuttosto simile e complessivamente equilibrata, nel campionamento rappresentativo le feature generate tendono ad addensarsi in modo evidente nella regione più densa del dominio reale (approssimativamente per valori di PC1 > 0 e PC2 < 0.2), lasciando invece parzialmente, o in alcuni casi completamente, non rappresentate le zone periferiche del dominio (in particolare per PC1 $<$ -0.3).
Questo comportamento spiega chiaramente il motivo per cui, nel dominio benigno, per il metodo rappresentativo la density risulta più elevata e la recall è più bassa rispetto agli altri tre.

3.3.4 Metriche con reti riaddestrate

Quando si utilizzano reti preaddestrate su ImageNet, le feature estratte rappresentano caratteristiche visive di tipo generale apprese su un ampio insieme di immagini naturali. In questo contesto, le metriche misurano la qualità percettiva globale delle immagini sintetiche, valutando quanto esse risultino realistiche in senso visivo generale, ma senza tenere conto delle specificità cliniche del dominio dermatologico. Diversamente, una rete riaddestrata (fine-tuned) su un dataset specifico di immagini di melanoma e nevo benigno apprende rappresentazioni più specializzate, legate a pattern pigmentari, bordi irregolari, variazioni cromatiche e simmetria della lesione.

L'analisi parallela delle due configurazioni consente di ottenere una valutazione più completa e bilanciata delle prestazioni del modello generativo. Le metriche calcolate con la rete preaddestrata forniscono una stima oggettiva della qualità percettiva generale, mentre quelle ottenute con la rete *fine-tuned* offrono una misura più specifica della qualità clinico-visiva delle immagini prodotte. Il confronto tra i due approcci risulta particolarmente informativo:

- una discrepanza tra i valori può indicare che il modello produca immagini realistiche, ma non perfettamente coerenti con i tratti distintivi delle lesioni;
- al contrario, una coerenza tra le due valutazioni suggerisce che il generatore sia in grado di coniugare realismo visivo e aderenza semantica al dominio medico.

FID e KID

Il calcolo delle metriche *FID* e *KID* mediante una rete *InceptionV3 riaddestrata* su immagini dermoscopiche produce rappresentazioni più coerenti con il dominio medico e riduce la distanza media tra immagini reali e sintetiche. Come previsto,

entrambe le metriche mostrano un miglioramento complessivo rispetto alle valutazioni ottenute con la rete pre-addestrata su ImageNet, evidenziando una maggiore sensibilità ai pattern clinici caratteristici delle lesioni cutanee.

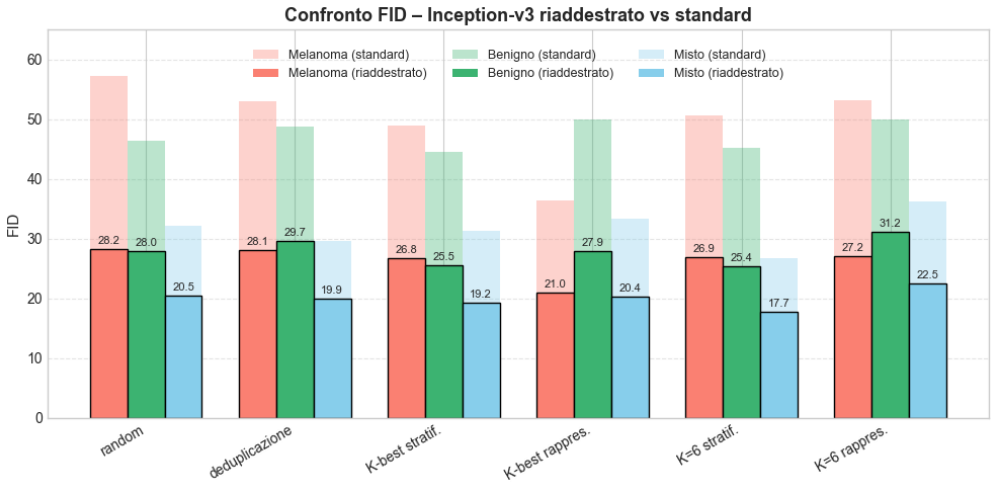


Figura 3.9: Valori di *FID* calcolati per i domini *melanoma* (in rosso), *benigno* (in azzurro) e *misto* (in verde) utilizzando l'estrattore di feature *Inception-v3 fine-tuned*. Le barre in ombra rappresentano, per ciascun dominio, i valori di FID ottenuti con la versione standard dell'Inception-v3, mostrati a scopo di confronto.

Tabella 3.3: Valori di *KID* calcolati con Inception-v3 riaddestrata per i domini melanoma, benigno e misto. I valori migliori per ciascun dominio sono evidenziati in grassetto.

Metodo	Melanoma	Benigno	Misto
Random	0.0101 ± 0.0007	**0.0067 ± 0.0004**	0.0045 ± 0.0004
Dedupl.	0.0105 ± 0.0008	0.0075 ± 0.0004	0.0043 ± 0.0004
K-best strat.	0.0094 ± 0.0007	0.0070 ± 0.0005	0.0045 ± 0.0004
K-best rappr.	**0.0049 ± 0.0003**	0.0093 ± 0.0005	0.0046 ± 0.0003
K=6 strat.	0.0099 ± 0.0007	0.0069 ± 0.0005	**0.0041 ± 0.0004**
K=6 rappr.	0.0093 ± 0.0006	0.0073 ± 0.0004	0.0048 ± 0.0003

Nel caso del *FID*, i valori si collocano tra 17.7 e 31.1, come riportato in Figura 3.9, indicando una qualità complessiva superiore e una maggiore coerenza con i benchmark riportati in letteratura. Per il *KID* (Tabella 3.3) i valori medi rientrano in un intervallo molto basso (tra 0.004 e 0.010) con deviazioni standard ridotte,

segno di una stima stabile e di una distribuzione statistica più coerente tra immagini reali e sintetiche.

Nel complesso, le osservazioni fatte per le metriche standard rimangono valide anche in questa configurazione: con la rete riaddestrata, le differenze tra le strategie di sottocampionamento risultano meno pronunciate, ma la gerarchia generale rimane invariata.

- Il campionamento *random* si conferma solido con risultati competitivi, mentre la *deduplicazione* non porta i miglioramenti attesi, mostrando prestazioni paragonabili o leggermente inferiori alla baseline.
- I metodi *stratificati* continuano a offrire delle ottime prestazioni globali, giustificate dal buon equilibrio che sussiste tra diversità e realismo.
- I metodi *rappresentativi*, in particolare *K-best rappresentativo*, confermano la loro efficacia nel dominio del melanoma, nonostante sia più ristretto e strutturato, a conferma che risultano particolarmente forti nel modellare realisticamente i pattern clinici ricorrenti; sono penalizzati nel dominio dei benigni a causa della limitata varietà generativa.

Un risultato interessante riguarda il rapporto tra i due domini: con la rete riaddestrata, le differenze tra melanoma e benigno si riducono sensibilmente. Il melanoma, che nella configurazione standard mostrava valori sistematicamente più elevati, presenta ora valori medi simili o persino inferiori rispetto al dominio benigno (21–28 contro 25–31).
La differenza tra le metriche ottenute con la rete InceptionV3 pre-addestrata su ImageNet e quelle calcolate con la versione riaddestrata su immagini dermoscopiche può essere interpretata alla luce della diversa natura dei due domini. La rete standard, ottimizzata per il riconoscimento di oggetti e scene naturali, fatica a rappresentare correttamente i melanomi, che presentano pattern visivi atipici e clinicamente rilevanti (bordi irregolari, pigmentazioni complesse e asimmetrie) non presenti nelle classi di ImageNet. La rete riaddestrata, invece, apprende features più clinicamente significative, migliorando la rappresentazione dei melanomi, che risultano più strutturati e coerenti. Questo si traduce in una riduzione del FID per il dominio del melanoma, poiché le immagini sintetiche vengono percepite come più simili a quelle reali.
Nel caso delle lesioni benigne, la situazione è diversa. Queste presentano forme semplici, simmetrie regolari e texture uniformi, caratteristiche visive più affini a quelle di oggetti naturali già presenti in ImageNet (come ombre e macchie). Di conseguenza, la rete pre-addestrata riesce già a rappresentarle in modo efficace e il riaddestramento non porta a un miglioramento significativo.

PRDC

Le metriche PRDC basate su una ResNet50 riaddestrata sul dominio dermoscopico dovrebbero produrre features più sensibili a dettagli clinici come pigmentazione, simmetria e irregolarità. In questo contesto, precision e recall assumono un significato più diagnostico, indicando quanto le immagini sintetiche rispettino le caratteristiche delle lesioni reali e quanto ne coprano la varietà.

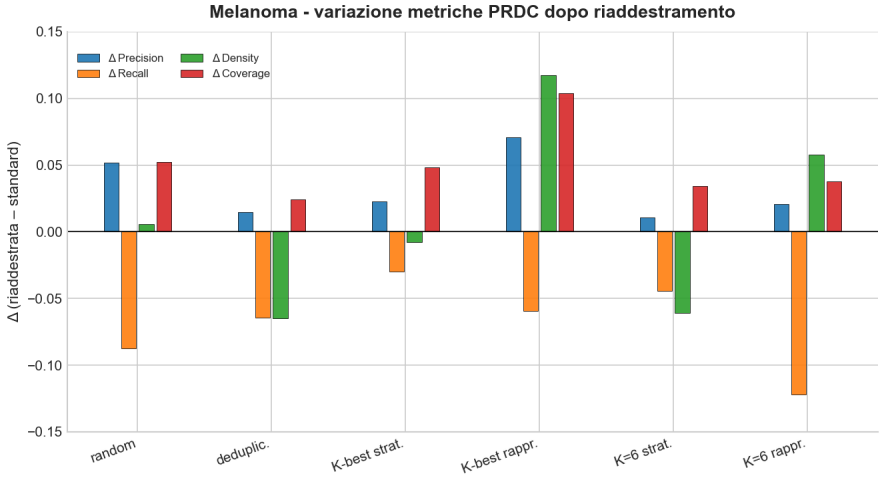
Metodo	**Melanoma**				**Benigno**			
	Prec.	Rec.	Dens.	Cov.	Prec.	Rec.	Dens.	Cov.
random	0.7870	0.4654	0.6443	0.5904	0.7994	**0.5524**	0.6880	0.6016
deduplic.	**0.8140**	0.5116	**0.7230**	0.6136	0.7726	0.5072	0.6621	0.6008
K-best strat.	0.7792	0.5508	0.6728	**0.6328**	0.7602	0.5042	0.6275	**0.5850**
K-best rappr.	0.7480	0.3410	0.6296	0.6100	0.8020	0.3540	**0.8705**	0.5524
K=6 strat.	0.7958	**0.5550**	0.6901	0.6222	**0.8088**	0.5344	0.6834	**0.6030**
K=6 rappr.	0.7472	0.3792	0.6514	0.5890	0.7428	0.3606	0.7336	0.5880

Tabella 3.4: Valori di *PRDC* calcolati con ResNet50 riaddestrata per i domini melanoma, benigno e misto. I valori migliori per ciascun dominio sono evidenziati in grassetto.

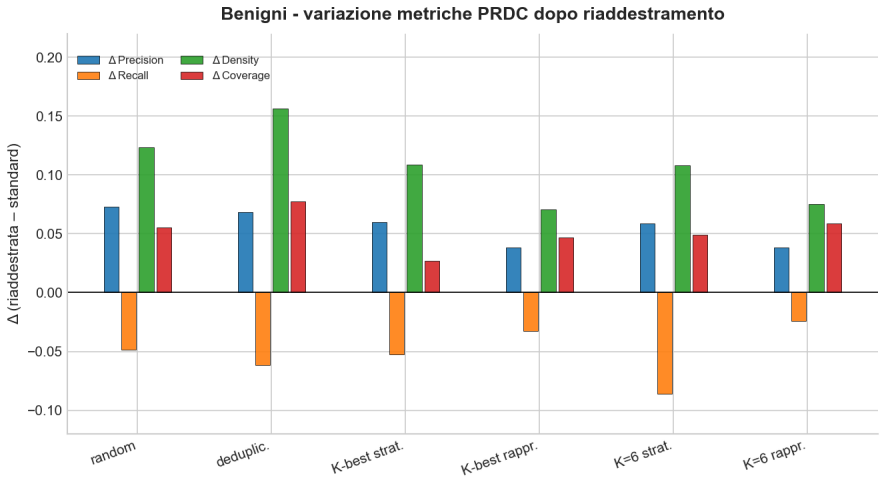
L'utilizzo della *ResNet50 riaddestrata* ha effettivamente migliorato la coerenza e la stabilità delle metriche PRDC, i cui nuovi valori sono riportati nella Tabella 3.4, evidenziando una maggiore sensibilità ai pattern clinici. Il riaddestramento ha anche ridotto la variabilità tra le strategie di campionamento, rendendo le differenze meno ampie ma più significative sul piano qualitativo. Tuttavia, questa specializzazione rende la metrica anche più selettiva, penalizzando immagini realistiche che non rientrano nei pattern appresi durante il fine-tuning. I valori medi di *Precision* e *Density* si collocano tendenzialmente in un intervallo compreso tra 0.74 e 0.87, mentre *Recall* e *Coverage* tra 0.38 e 0.71, confermando un buon equilibrio tra qualità visiva e varietà delle immagini sintetiche.

L'impiego della ResNet50 riaddestrata ha portato a variazioni coerenti con la maggiore specializzazione clinica del modello, visualizzabili nelle Figure 3.10a e 3.10b:

- La *precision* migliora in entrambi i domini e per tutti i metodi, segno che le immagini sintetiche vengono riconosciute più spesso come realistiche rispetto alle immagini reali. Questo è probabilmente dovuto al fatto che la rete, essendo addestrata su lesioni dermoscopiche, è più brava a cogliere i dettagli clinici rilevanti.



(a) Dominio melanoma



(b) Dominio benigno

Figura 3.10: Variazioni delle metriche PRDC tra valori calcolati con feature estratte da modelli fine-tuned e quelli standard, nei domini *melanoma* e *benigno*.

- La *density* aumenta in tutti i casi nel dominio benigno, mentre nei melanomi si riduce leggermente, fatta eccezione per i campionamenti cluster-based rappresentativi. Questo può significare che nei benigni la maggiore varietà visiva consente alla rete di riconoscere più facilmente sottostrutture cliniche, premiando le immagini sintetiche che vi si collocano. Nei melanomi, invece, la rete diventa più selettiva, e solo le immagini che rispettano rigorosamente i pattern appresi vengono riconosciute come appartenenti a regioni dense (come

nel caso del metodo rappresentativo), riducendo così il valore della metrica.

- La *recall* peggiora in entrambi i casi, probabilmente a causa della maggiore selettività: la rete riconosce con maggiore sicurezza le immagini che rientrano nei pattern appresi, ma penalizza quelle che si discostano, anche se realistiche, portando a una copertura più ristretta del dominio reale.

- La *coverage*, infine, mostra valori più stabili e leggermente superiori in tutti i casi. Questo suggerisce che, pur con una riduzione della recall, la rete riesce comunque a mantenere una buona rappresentazione complessiva del dominio, grazie a una maggiore robustezza nella definizione delle regioni semantiche.

Confrontando le varie strategie di raffinamento del set di training, i risultati PRDC con ResNet50 riaddestrata confermano tendenze già osservate con il feature extractor standard.
Le strategie *stratificate* si distinguono per equilibrio e costanza, con i valori medi più alti di *Recall* (0.53–0.59) e tra i migliori in *Precision* (0.79–0.81), dimostrando una buona copertura del dominio senza compromettere la coerenza visiva. Anche la *deduplicazione* mostra prestazioni solide, simili a quelle ottenute per la random, ma con precision più elevata a conferma dell'effetto positivo della deduplicazione. Le strategie *rappresentative*, invece, penalizzano recall e coverage (0.34–0.43), ma ottengono density nettamente superiore, soprattutto nel dominio benigno (es. 0.8785 per K-best), che riflette una generazione più centrata e pulita, a scapito della varietà.

3.4 Classificatori

Valutiamo ora le prestazioni dei classificatori parzialmente riaddestrati *InceptionV3* e *ResNet50* sulle immagini sintetiche, confrontando le rispettive accuracy e formulando ipotesi interpretative sui risultati ottenuti.
Seguirà l'analisi basata sul classificatore vincitore della *ISIC Challenge 2020*, considerato tra i più affidabili nel contesto dermoscopico.

3.4.1 Inception-v3 e ResNet50

L'accuracy di *InceptionV3* parzialmente riaddestrato sulle immagini dermoscopiche reali si attesta al 67%, un valore che riflette una competenza moderata e sottolinea la complessità del task: le lesioni dermoscopiche presentano elevata variabilità intra-classe, con numerosi casi borderline e sovrapposizioni semantiche.
Sulle immagini sintetiche, invece, l'accuracy varia sensibilmente in base al metodo di campionamento, come è osservabile dai risultati riportati nella Tabella 3.5. I *metodi rappresentativi* (soprattutto *K-best*) raggiungono valori più alti, fino a circa 0.77, mentre le strategie stratificate, random e deduplicazione si fermano tra 0.52 e 0.60. Questo divario è indicativo: i metodi che mantengono maggiore varietà, come stratificato e random, generano immagini che riflettono meglio la distribuzione reale, inclusi i casi di confine, rendendo il compito più difficile per il classificatore, che mostra un'accuratezza inferiore. Al contrario, le immagini generate con approcci centrati sono più "pulite", regolari e vicine ai prototipi di classe, quindi più facilmente riconoscibili dal classificatore. In pratica, il generatore, guidato da campioni rappresentativi, tende a rafforzare le caratteristiche diagnostiche chiave, riproducendole in modo più fedele e creando quindi una sinergia con il classificatore. Tuttavia, un'accuracy superiore a quella ottenuta sul set reale spesso riflette un set sintetico più semplice e meno vario. Questa osservazione per i metodi di sottocampionamento rappresentativi è coerente con i risultati delle metriche FID/KID, che rimarcano il buon allineamento globale delle immagini generate, e PRDC, che rivela una "density" molto elevata a discapito di una copertura più bassa, a conferma di come tali metodi privilegino la qualità visiva a scapito della diversità. Si osserva inoltre che i risultati ottenuti in termini di *precision* e *recall*, sia complessivamente sia per i singoli domini, mostrano gli stessi andamenti e lo stesso ordine di prestazioni riscontrati per l'*accuracy*.

L'adozione della *ResNet50 riaddestrata* ha portato a risultati più coerenti con la complessità del dominio dermoscopico, migliorando l'accuracy sul validation set reale al 78% e confermando che le feature apprese dalla ResNet50 sono più adatte a riconoscere strutture cliniche rilevanti.

Metodo	Accur.	Melanoma		Benigno	
		Prec.	Rec.	Prec.	Rec.
K-best rappresent.	0.76	0.77	0.75	0.76	0.77
K=6 rappresent.	0.68	0.67	0.70	0.69	0.66
Reali	**0.67**	0.64	0.76	0.71	0.58
K-best stratificato	0.60	0.59	0.64	0.61	0.56
random	0.54	0.53	0.57	0.54	0.50
K=6 stratificato	0.53	0.53	0.57	0.53	0.49
deduplicazione	0.52	0.52	0.57	0.52	0.47

Tabella 3.5: Risultati classificazione con *Inception-v3* sulle immagini generate per i diversi metodi di campionamento. Sono riportati Accuracy, Precision e Recall distinti per classi.

Per quanto riguarda i test sulle immagini sintetiche, i cui risultati sono riportati nella Tabella 3.6, le strategie di raffinamento del set di training *stratificate* e *randomiche*, pur garantendo maggiore varietà, ottengono valori leggermente inferiori di accuracy rispetto ai metodi *rappresentativi*, che si confermano i più efficaci, con *K-best* che raggiunge 0.86.

L'aumento complessivo delle performance sul set reale è accompagnato da un ampliamento del divario tra i diversi metodi: la rete enfatizza la qualità strutturale e premia le immagini che riproducono con maggiore fedeltà i pattern diagnostici canonici. Inoltre, sembra essere più sensibile a texture e dettagli clinici, tendendo a penalizzare le immagini affette da rumore, artefatti o strutture ibride, che sono più frequenti nei campionamenti eterogenei. Anche in questo caso le migliori prestazioni dei metodi basati sui centroidi si spiegano con il fatto che le immagini generate riproducono più accuratamente le caratteristiche cliniche dominanti, risultando più facilmente riconoscibili dal classificatore.

Il confronto tra InceptionV3 e ResNet50, entrambe riaddestrate su immagini dermoscopiche, mostra come l'architettura influenzi la percezione della qualità delle immagini sintetiche. InceptionV3 mantiene una sensibilità più generale e meno selettiva, valutando positivamente anche immagini realistiche ma non perfettamente aderenti ai prototipi clinici, mentre ResNet50, invece, mostra una maggiore capacità discriminativa. Tra i metodi di raffinamento utilizzati per selezionare il set di training della rete generativa, quello *cluster-based rappresentativo* con *"K-best"* ha prodotto risultati persino superiori a quelli ottenuti con immagini reali, suggerendo che il generatore ha appreso con efficacia le *features diagnostiche distintive* più rilevanti, escludendo tuttavia i campioni meno centrali dei due domini.

Metodo	Accur.	Melanoma		Benigno	
		Prec.	Rec.	Prec.	Rec.
K-bests rappresent.	0.86	0.89	0.83	0.84	0.89
K=6 rappresent.	0.78	0.80	0.74	0.76	0.82
Reali	**0.78**	0.75	0.86	0.83	0.71
K-bests stratificato	0.72	0.73	0.68	0.70	0.75
deduplicazione	0.68	0.70	0.64	0.67	0.72
random	0.66	0.66	0.63	0.65	0.68
K=6 stratificato	0.64	0.65	0.64	0.64	0.65

Tabella 3.6: Risultati classificazione con *ResNet50* sulle immagini generate per i diversi metodi di campionamento. Sono riportati Accuracy, Precision e Recall distinti per classi.

3.4.2 Classificatore vincitore della ISIC Challenge 2020

Il classificatore vincitore della ISIC Challenge 2020 rappresenta un riferimento clinicamente più significativo rispetto alla semplice accuracy, in quanto valuta la capacità del modello di discriminare correttamente tra casi positivi e negativi su un ampio spettro di soglie decisionali. La *metrica AUC (Area Under the Curve)* consente di misurare la qualità discriminativa di un classificatore in modo robusto, soprattutto in contesti clinici dove è fondamentale bilanciare sensibilità e specificità. L'AUC rappresenta l'area sotto la curva ROC (Receiver Operating Characteristic) e quantifica la probabilità che il modello assegni un punteggio più alto a un campione positivo rispetto a uno negativo: valori più vicini a 1 indicano una maggiore capacità del classificatore di distinguere correttamente tra classi.

Secondo quanto riportato da [59], sul validation set reale il classificatore raggiunge un'AUC di 0.9490, a conferma dell'efficacia del modello nel riconoscere strutture cliniche rilevanti.
Sulle immagini sintetiche, invece, si osserva un calo progressivo dell'AUC in funzione della strategia di campionamento adottata per il training della rete generativa. I valori, riportati nella Tabella 3.7, spaziano da 0.90 nei metodi rappresentativi a circa 0.60 nei metodi stratificati e randomici, confermando i risultati riportati dagli altri classificatori ed evidenziando come la qualità diagnostica delle immagini sintetiche dipenda fortemente dalla composizione del dataset di partenza.

I metodi *K-best rappresentativo* e *K-means 6 rappresentativo* ottengono le AUC più elevate (0.9048 e 0.8973), molto vicine a quelle ottenute su immagini reali, confermando che le immagini generate in questi casi preservano efficacemente le features discriminative apprese dal classificatore.
I metodi *stratificati* e *randomici* mostrano invece AUC più basse (tra 0.60 e 0.77),

Metodo	AUC
Reali	**0.9490**
K-best rappresentativo	0.9048
K=6 rappresentativo	0.8973
K-best stratificato	0.7729
deduplicazione	0.6504
random	0.6169
K=6 stratificato	0.5991

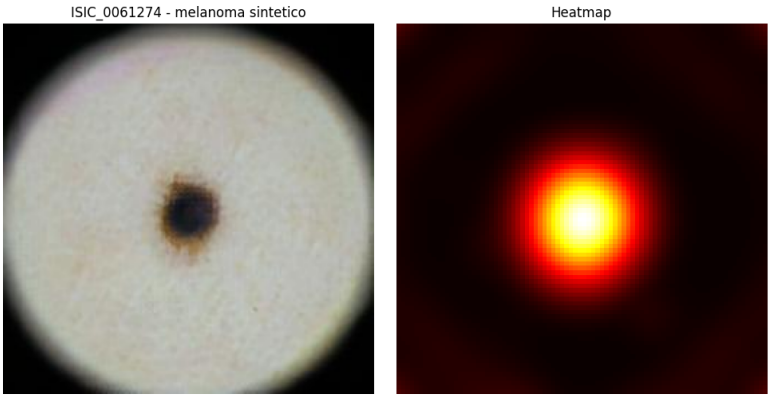
Tabella 3.7: Risultati del classificatore *vincitore su ISIC 2020*. Il valore di AUC è riportato per ciascun metodo di campionamento.

ribadendo che nonostante una migliore copertura distribuzionale, le immagini generate risultano morfologicamente meno diagnostiche.

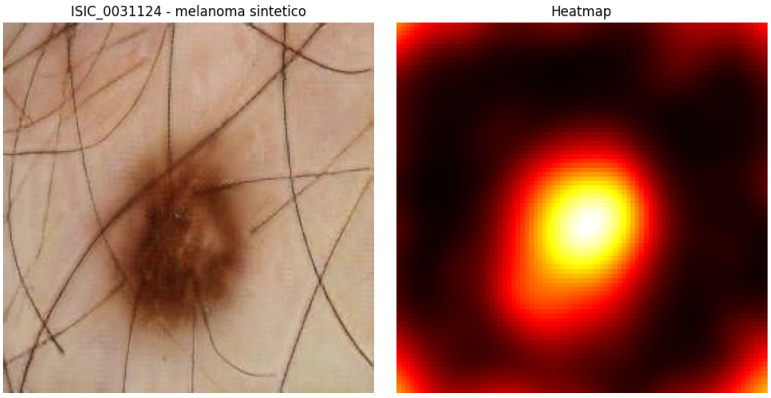
In conclusione, i risultati ottenuti dal classificatore vincitore della ISIC Challenge 2020 non fanno altro che rafforzare la validità delle osservazioni e delle ipotesi formulate nelle sezioni precedenti. Il fatto che tali tendenze emergano in modo coerente anche con un modello altamente performante e clinicamente affidabile conferma la solidità delle analisi condotte e l'efficacia delle strategie di valutazione adottate.
Inoltre, l'andamento della AUC è perfettamente coerente con quanto osservato nelle metriche FID, KID e PRDC:

- **Metodi rappresentativi**: massimizzano la fedeltà semantica, generando immagini realistiche e coerenti con le classi diagnostiche.
- **Metodi stratificati e randomici**: ottimizzano la diversità statistica, ma a scapito della chiarezza strutturale e del realismo clinico.
- **Metodo basato su deduplicazione con cosine similarity**: nel complesso non ha portato a miglioramenti rilevanti rispetto alla selezione casuale dei campioni di training.

Infine, viene mostrata in Figura 3.11 la *heatmap di attivazione* generata dal classificatore vincitore della challenge ISIC 2020 per due lesioni di melanomi sintetici. La mappa evidenzia le regioni dell'immagine che hanno maggiormente contribuito alla classificazione, con intensità crescente dal nero (bassa attivazione) al bianco (massima attivazione).
Si osserva una forte attivazione centrata sulle lesioni pigmentate, coerente con l'interpretazione clinica e con il comportamento atteso da un classificatore ben



(a) ISIC_0061274



(b) ISIC_0031124

Figura 3.11: A sinistra, l'immagine dermatoscopica sintetica di melanoma passata come input; a destra la heatmap di attivazione generata dal classificatore vincitore della challenge ISIC 2020 per tale lesione: più il colore tende verso il bianco, maggiormente tale zona è stata presa in considerazione.

addestrato. Le aree periferiche e i dettagli non rilevanti, come i peli, i bordi scuri e la pelle circostante, risultano correttamente ignorati. Questo tipo di visualizzazione consente di valutare la trasparenza e la localizzazione semantica del modello, offrendo un primo livello di interpretabilità utile sia in ambito clinico che computazionale.

3.5 Valutazione visiva con criteri ABCDE

Dopo aver analizzato le performance quantitative del modello nelle diverse configurazioni, proponiamo ora un'analisi qualitativa visiva basata sui criteri ABCDE. Nella Figura 3.12 sono riportati sei esempi di nevi benigni reali (prima colonna) e le rispettive immagini sintetiche generate a partire dai diversi metodi di sottocampionamento. Nella Figura 3.13, analogamente, sono mostrati sei melanomi reali seguiti dalle immagini generate per ciascun metodo.

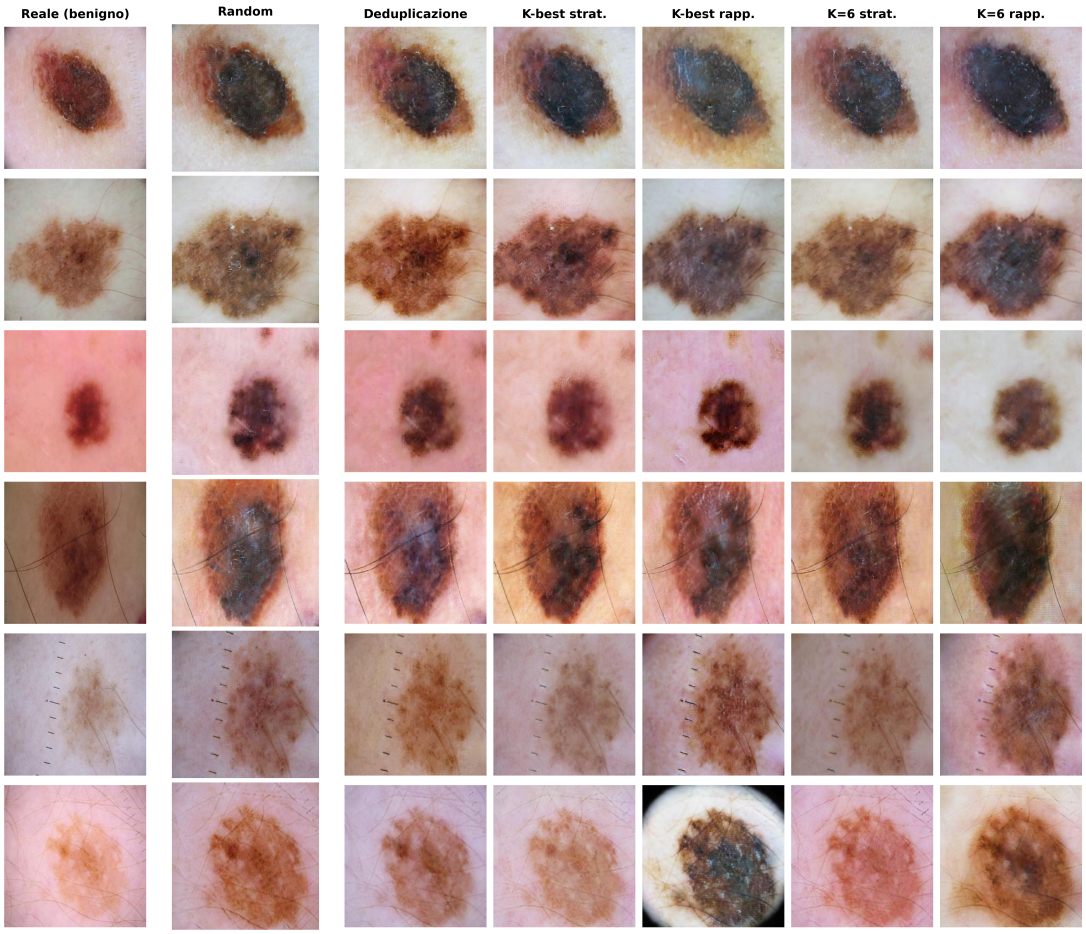


Figura 3.12: Confronto qualitativo tra immagini reali di nevi benigni (prima colonna a sinistra) e corrispondenti *melanomi generati* a partire da diversi metodi di sottocampionamento (nelle colonne successive); nell'ordine: *random*, *deduplicazione*, *K-best stratificato*, *K-best rappresentativo*, *K=6 stratificato*, *K=6 rappresentativo*.

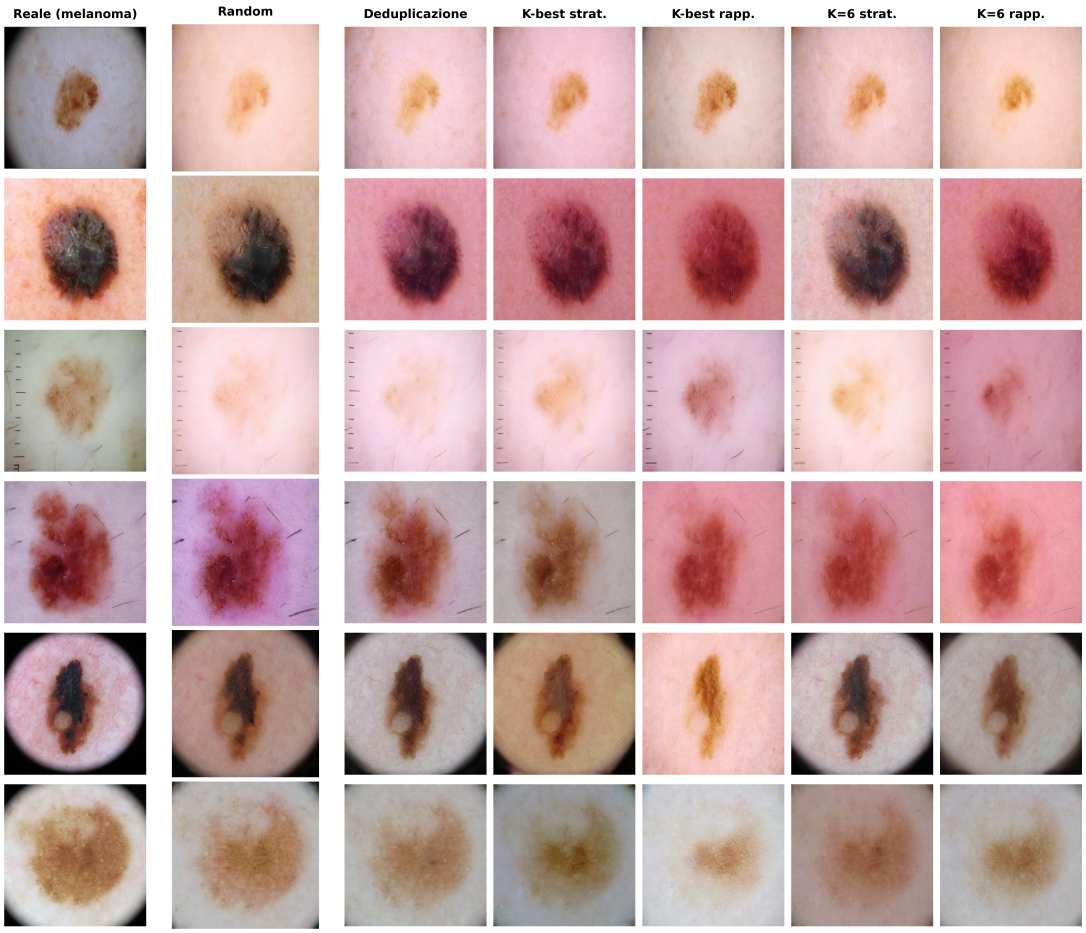


Figura 3.13: Confronto qualitativo tra immagini reali di melanomi reali (prima colonna a sinistra) e corrispondenti *nevi benigni generati* a partire da diversi metodi di sottocampionamento (nelle colonne successive); nell'ordine: *random*, *deduplicazione*, *K-best stratificato*, *K-best rappresentativo*, *K=6 stratificato*, *K=6 rappresentativo*.

Nel complesso, le immagini sintetiche risultano visivamente coerenti e di buona qualità. Si osserva, in particolare, una tendenza alla scuritura delle lesioni nei melanomi e alla schiaritura nei nevi, in linea con le caratteristiche cliniche attese. Anche la dimensione delle lesioni sembra variare in modo plausibile: alcuni melanomi appaiono espansi rispetto alla lesione benigna di partenza, mentre alcuni nevi risultano più contenuti rispetto ai melanomi originari.
Tuttavia, l'asimmetria non viene significativamente alterata: la forma della lesione tende a rimanere simile all'originale, suggerendo che la CycleGAN, pur efficace nel trasferimento di stile, fatichi a modificare la struttura morfologica profonda

dell'immagine. È interessante notare come tutte le configurazioni riescano a preservare dettagli estranei alla lesione, come peli e tacche di misurazione, segno di una buona capacità di ricostruzione contestuale. In alcuni casi, il generatore modifica anche il colore della pelle circostante, indicando la necessità di una futura normalizzazione cromatica per analisi più precise.

Osservando i melanomi sintetici, tutti i metodi tendono a scurire la lesione, ma i metodi cluster-based rappresentativi producono tonalità più intense e marcate. Questo potrebbe spiegare i risultati più convincenti ottenuti con questi campioni e il fatto che i classificatori abbiano riconosciuto più facilmente tali immagini come melanomi. Un esempio emblematico è l'ultima immagine generata con il metodo *K-best rappresentativo*, che ha una tonalità nettamente più simile a quella standard di un melanoma, anche se non riesce a riprodurre le tacche di misurazione e introduce il bordo del campo visivo del dermoscopio, probabilmente appreso da immagini reali presenti nel dataset.

Per quanto riguarda le lesioni benigne sintetiche, il metodo *K-best rappresentativo* tende a schiarire significativamente il colore e a ridurre l'estensione della lesione, rendendola visivamente più simile a un nevo benigno. Tuttavia, in alcuni casi si perde la struttura originale dell'immagine, come il cono visivo, mentre altre configurazioni, più fedeli, mantengono il colore più vicino all'originale, con il rischio di confondere il classificatore.

Queste osservazioni aiutano a interpretare i risultati ottenuti: il *metodo rappresentativo* sembra confermare una maggiore coerenza semantica a discapito di una variabilità interna ridotta, con immagini simili a quelle prototipiche, tendenzialmente tutte scure (melanomi) o tutte chiare (nevi), che però agevolano in particolare i classificatori. Al contrario, configurazioni come quella stratificata sembrano aver appreso anche caratteristiche da lesioni meno comuni, contribuendo a una maggiore diversità e a una rappresentazione più equilibrata del dataset.

3.6 Ottimizzazione

Sulla base delle evidenze raccolte, si è giunti alla conclusione che il metodo di sottocampionamento *cluster-based rappresentativo*, con valori di k determinati separatamente per i due domini tramite *silhouette score* e *Davies–Bouldin index*, rappresenta la strategia di raffinamento più promettente per la costruzione del dataset di training, nonostante la limitata copertura della distribuzione reale.

La fase di ottimizzazione finale per migliorare ulteriormente la qualità delle immagini sintetiche, di conseguenza, mira a enfatizzare i punti di forza del metodo rappresentativo, ossia la capacità di catturare pattern diagnostici centrali, e al tempo stesso a mitigare la ridotta copertura del dominio reale. È sulla base di queste motivazioni

che il nuovo addestramento combina un sottocampionamento del set di training *cluster-based rappresentativo* (K-means con k selezionato tramite *silhouette score* e *Davies–Bouldin index*), un'estensione fino a 500 epoche, in linea con quanto proposto da [29], e un incremento del numero di immagini di addestramento per ciascun dominio. Questa configurazione di training si è dimostrata efficace sotto molteplici aspetti: le curve di *loss* mostrano una convergenza stabile, con la generator loss che raggiunge valori significativamente più bassi rispetto alla configurazione iniziale (Figura 3.14), mentre i valori medi di output del discriminatore per le immagini reali e sintetiche di melanoma ("M_{real}" ≈ 0.6 e "M_{fake}" ≈ 0.4) si avvicinano molto al limite teorico di 0.5, con un andamento regolare e privo di oscillazioni (Figura 3.15).

Le metriche di valutazione confermano il miglioramento complessivo: osservando i valori riportati nella Tabella 3.8, FID e KID risultano i più bassi tra tutte le configurazioni testate e, parallelamente, quelli di precision e density mantengono valori elevati, indicando un'elevata qualità globale delle immagini generate e fedeltà delle caratteristiche clinicamente rilevanti. Viene registrato inoltre un incremento significativo di recall e coverage, che non penalizza le altre metriche, suggerendo che il modello ha migliorato la copertura del manifold reale, preservando al contempo una buona coerenza visiva e diagnostica delle immagini sintetiche. Per quanto riguarda i classificatori esterni, i risultati nella Tabella 3.9 mostrano che le prestazioni si mantengono elevate, anche se non si osservano miglioramenti significativi rispetto alla configurazione iniziale: si registra una lieve riduzione dell'accuratezza per entrambi i modelli (Inception-v3 e ResNet-50), mentre il valore di AUC relativo al più affidabile classificatore vincitore della challenge ISIC 2020 resta pressoché invariato, a conferma della stabilità del comportamento discriminativo.

Infine, sebbene una valutazione clinica definitiva richiederebbe il parere di un dermatologo, le immagini generate appaiono visivamente realistiche e coerenti con le caratteristiche attese: come si può osservare dalle Figure 3.16 e 3.17, l'analisi visiva conferma la corretta separazione dei due domini, con le lesioni melanomiche che presentano tonalità più scure e forme espanse, coerenti con i criteri diagnostici della regola ABCDE, mentre i nevi benigni risultano invece più piccoli, regolari e omogenei nelle tonalità. Inoltre, si può osservare che anche gli elementi estranei alla lesione, come peli o bordi scuri, vengono riprodotti dal generatore con buona fedeltà. Infine, è possibile osservare l'evoluzione di alcuni nevi benigni reali in melanomi sintetici (Figura 3.18) e la ricostruzione del nevo originale sintetico da cui plausibilmente può essersi formato il melanoma reale (Figura 3.19).

In conclusione, questa strategia di training ha portato alla generazione di immagini sintetiche ancor più realistiche e, soprattutto, clinicamente rilevanti con una migliore copertura del dominio reale, come confermato dai risultati ottenuti attraverso il nostro sistema di validazione.

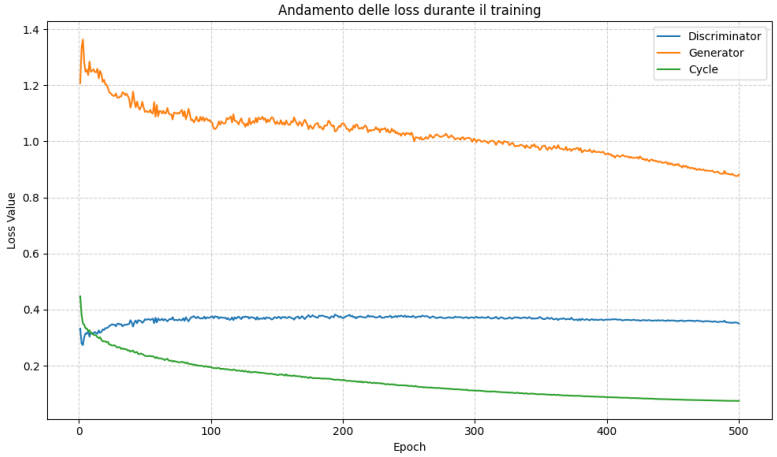


Figura 3.14: Andamento della training loss per addestramento con configurazione ottimizzata. In arancione la loss del generatore, in blu quella del discriminatore e in verde la cycle-consistency loss.

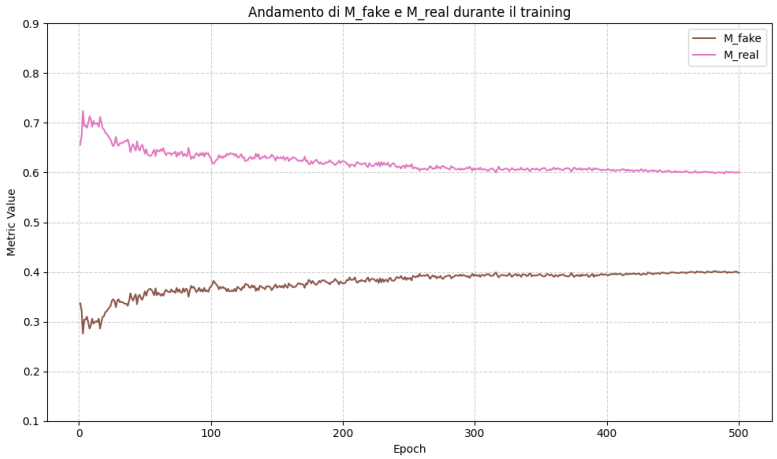


Figura 3.15: Andamento dell'output del discriminatore sulle immagini di melanoma reali ("M_{real}" - in rosa) e sintetiche ("M_{fake}" - in marrone) per addestramento con configurazione ottimizzata.

Tabella 3.8: Valori delle metriche di valutazione *FID*, *KID* e *PRDC* calcolati sui tre domini dopo l'addestramento del modello con *sottocampionamento cluster-based rappresentativo* basato su *K-means* (K=3 per il dominio melanoma e K=4 per il dominio benigno), esteso a 500 epoche e con 4000 immagini di training.

Metrica	**Melanoma**	**Benigno**	**Misto**
FID			
Standard	31.27	42.46	29.81
Riaddestrata	15.61	21.01	15.58
KID			
Standard	0.0133 ± 0.0012	0.0224 ± 0.0013	0.0120 ± 0.0010
Riaddestrata	0.0044 ± 0.0006	0.0048 ± 0.0004	0.0032 ± 0.0004
PRDC			
Precision	0.769	0.760	0.777
Recall	0.546	0.415	0.563
Density	0.658	0.784	0.774
Coverage	0.671	0.574	0.706
PRDC (ResNet50 riaddestrata)			
Precision	0.801	0.789	0.817
Recall	0.512	0.432	0.557
Density	0.747	0.786	0.810
Coverage	0.711	0.608	0.723

Tabella 3.9: Prestazioni dei classificatori esterni sulle immagini sintetiche generate dal modello addestrato con *sottocampionamento cluster-based rappresentativo* basato su *K-means* (K=3 per il dominio melanoma e K=4 per il dominio benigno), esteso a 500 epoche e con 4000 immagini di training.

Modello	**Prestazione**
Inception-V3	Accuracy = 71%
ResNet50	Accuracy = 83%
Vincitore ISIC 2020	AUC = 0.9039

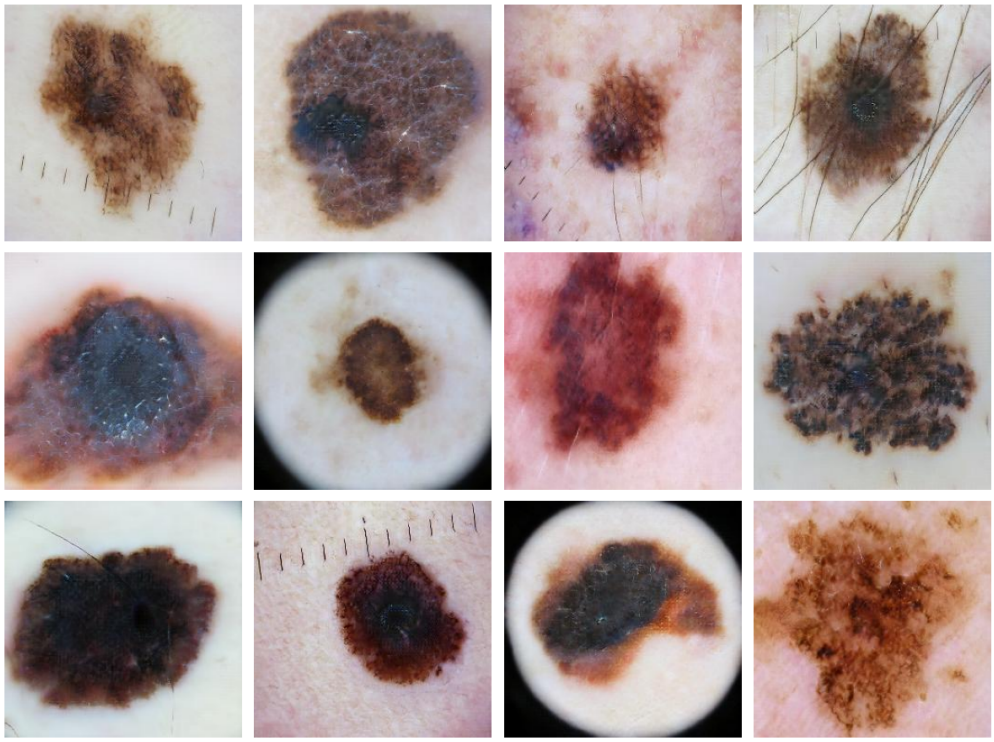


Figura 3.16: Esempi di immagini sintetiche di *melanoma* generate dal modello addestrato con *sottocampionamento cluster-based rappresentativo* basato su *K-means* (K=3 per il dominio melanoma e K=4 per il dominio benigno), esteso a 500 epoche e con 4000 immagini di training.

Figura 3.17: Esempi di immagini sintetiche di *nevi benigni* generate dal modello addestrato con *sottocampionamento cluster-based rappresentativo* basato su *K-means* (K=3 per il dominio melanoma e K=4 per il dominio benigno), esteso a 500 epoche e con 4000 immagini di training.

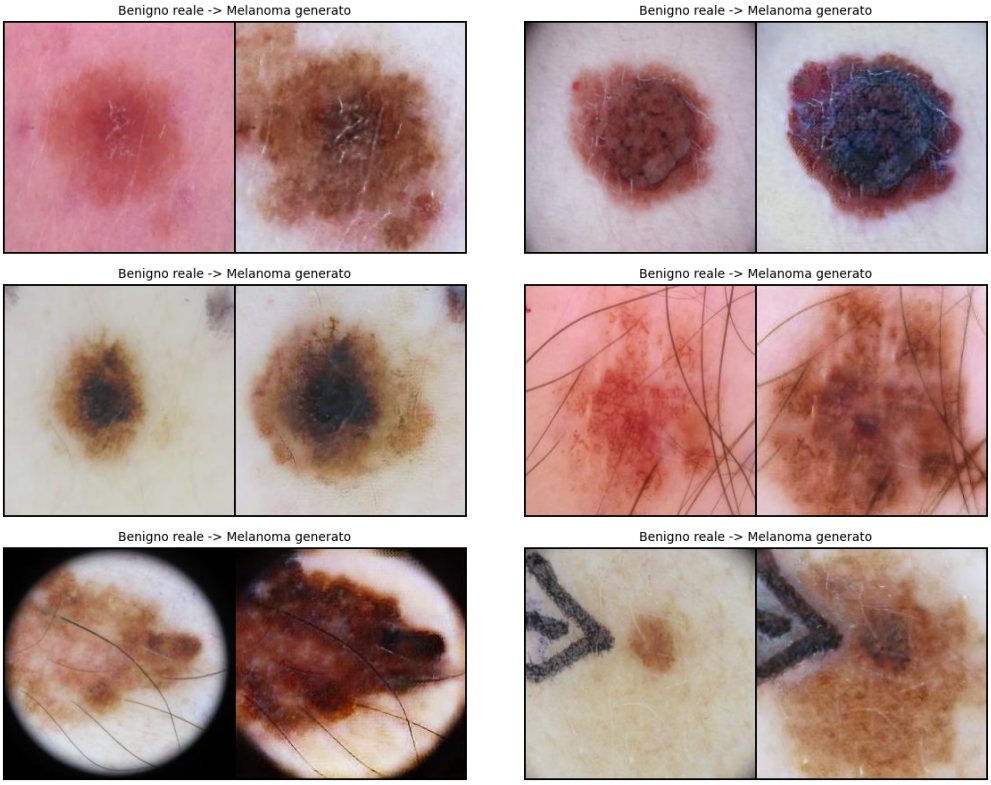


Figura 3.18: Confronto visivo tra un nevo reale (a sinistra) e la sua plausibile evoluzione sintetica in melanoma generata dal modello addestrato con *sottocampionamento cluster-based rappresentativo* basato su *K-means* (K=3 per il dominio melanoma e K=4 per il dominio benigno), esteso a 500 epoche e con 4000 immagini di training (a destra).

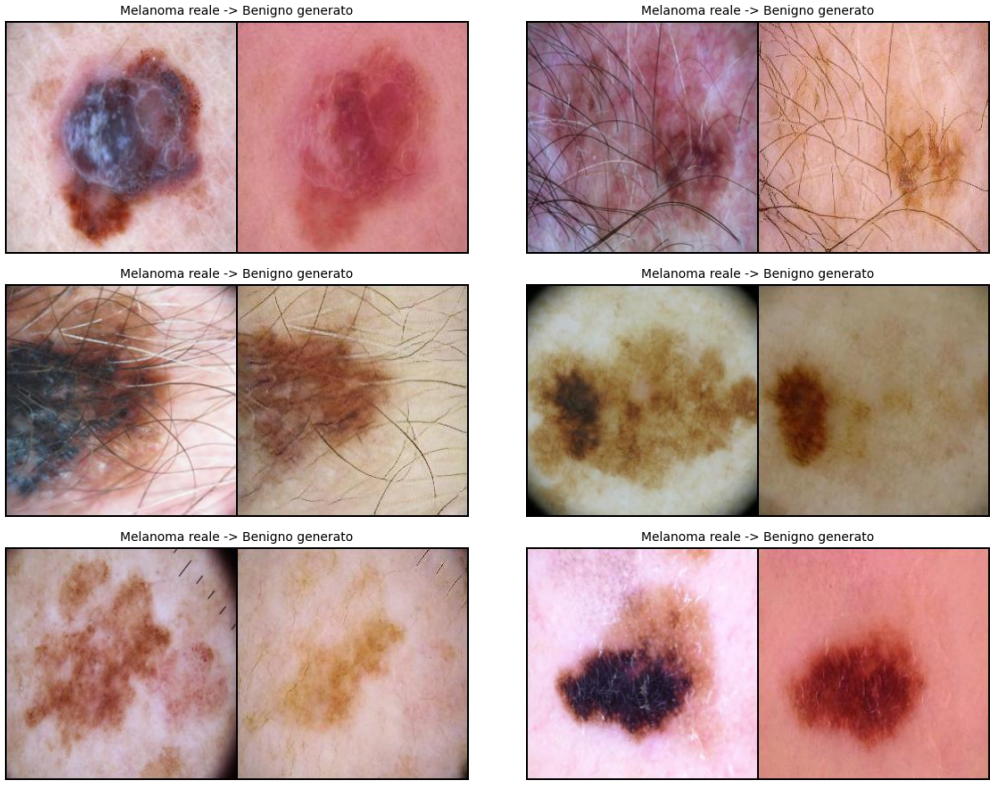


Figura 3.19: Confronto visivo tra un melanoma reale (a sinistra) e la plausibile ricostruzione sintetica del neo benigno precursore generata dal modello addestrato con *sottocampionamento cluster-based rappresentativo* basato su *K-means* (K=3 per il dominio melanoma e K=4 per il dominio benigno), esteso a 500 epoche e con 4000 immagini di training (a destra).

Capitolo 4

Conclusioni

Questa tesi si è posta come obiettivo principale la generazione di immagini dermoscopiche sintetiche di melanoma attraverso l'impiego di reti *Generative Adversarial Network (GAN)*.
L'esplorazione della letteratura sulle diverse architetture di GAN ha guidato la scelta della *CycleGAN*, selezionata per la sua capacità di apprendere trasformazioni bidirezionali tra domini visivi. Questa caratteristica si è rivelata particolarmente interessante nel contesto dermoscopico, permettendo da un lato di simulare l'evoluzione di un nevo benigno in melanoma e dall'altro di ricostruire un possibile aspetto originario del nevo a partire da un'immagine di melanoma reale.
Il lavoro si è sviluppato prevalentemente con un approccio metodologico e sperimentale, incentrato sulla valutazione comparativa dei modelli generativi e sull'analisi sistematica di diverse strategie di sottocampionamento del dataset di training.

In una fase preliminare, l'attenzione è stata rivolta all'analisi dei dataset reali utilizzati per l'addestramento e la validazione dei modelli generativi. La scelta è ricaduta sui dataset *ISIC 2019* e *ISIC 2020*, ampiamente riconosciuti nella comunità scientifica per la qualità delle immagini e l'affidabilità clinica delle annotazioni.
Lo studio approfondito dei metadati associati alle immagini ha fornito indicazioni utili sulla distribuzione e la composizione del campione, evidenziando al contempo alcune criticità intrinseche, come lo sbilanciamento tra classi e la presenza di immagini duplicate o molto simili, osservazioni che hanno orientato la successiva fase di sperimentazione verso la scelta di strategie di campionamento mirate. L'unione dei due insiemi ha permesso di ottenere un database più esteso e bilanciato con un aumento complessivo del numero di immagini di melanoma, categoria notoriamente meno rappresentata rispetto alle lesioni benigne. È seguita una fase di scrematura del dataset unificato sulla base di criteri coerenti con la letteratura scientifica relativa allo sviluppo del melanoma e adattati al task generativo specifico.

Dopo una prima fase di sperimentazione e di generazione delle immagini mediante

la struttura architetturale della *CycleGAN* proposta da Zhu et al. (2017) [22] e successivamente ripresa da Jutte et al. (2024) [1], si è posto il problema della valutazione delle immagini prodotte. A partire da questa esigenza, il lavoro si è concentrato sullo sviluppo di un metodo di validazione completo e strutturato, in grado di misurare in modo oggettivo sia il realismo percettivo sia la fedeltà clinica delle immagini sintetiche rispetto alle corrispondenti reali.
In primo luogo, è stata condotta una valutazione interna del training, basata sull'andamento delle funzioni di *loss*, del FID e sugli *output medi dei discriminatori* ("M_{real}" e "M_{fake}") nel dominio del melanoma, monitorati al variare delle epoche per valutare la stabilità e la convergenza del modello.
A questa analisi si è affiancata una valutazione esterna attraverso le principali metriche quantitative utilizzate in letteratura per i modelli GAN, ovvero *Fréchet Inception Distance (FID)*, *Kernel Inception Distance (KID)* e *PRDC* (Precision, Recall, Density, Coverage), calcolate sia nella loro *versione standard*, basata su reti pre-addestrate su ImageNet, sia utilizzando reti parzialmente riaddestrate su immagini dermoscopiche, al fine di ottenere una misura più sensibile alle caratteristiche cliniche proprie del dominio. Queste misure hanno consentito di stimare la somiglianza statistica tra immagini reali e sintetiche, la copertura della distribuzione reale e la capacità del generatore di mantenere un equilibrio tra realismo visivo e varietà interna.
Infine, per completare il quadro valutativo, sono stati utilizzati classificatori supervisionati con l'obiettivo di verificare in che misura il generatore avesse effettivamente appreso e riprodotto le feature diagnostiche rappresentative delle due classi di lesione.

L'analisi ha messo inoltre in evidenza l'importanza cruciale della fase di selezione e costruzione del set di training, che si è rivelata determinante per la qualità finale delle immagini generate proprio grazie all'applicazione del metodo di validazione sviluppato. Sono state testate diverse strategie di sottocampionamento, in particolare *randomica*, *deduplicata*, *stratificata* e *rappresentativa*, ciascuna caratterizzata da un diverso equilibrio tra varietà, compattezza e coerenza interna del dataset.

Nel complesso, i risultati ottenuti dimostrano che la *CycleGAN*, opportunamente addestrata su un dataset bilanciato e costruito con criteri clinicamente motivati, è in grado di generare immagini dermoscopiche sintetiche realistiche, coerenti e utili ai fini dell'analisi automatizzata. Inoltre, le tecniche di valutazione adattate confermano che la composizione del training set influisce in modo diretto sulla stabilità e sull'efficacia dell'apprendimento della rete generativa. In particolare, i metodi *cluster-based stratificati* si sono distinti per una maggiore stabilità del training e una rappresentatività più equilibrata del dominio, garantendo una copertura più ampia delle varianti cliniche. Al contrario, le strategie *cluster-based rappresentative*, basate sui centroidi dei cluster, hanno prodotto immagini più

coerenti e realistiche, con un andamento degli output discriminatori più vicino alle aspettative teoriche. Questi metodi hanno permesso al generatore di apprendere con maggiore fedeltà le caratteristiche diagnostiche salienti, pur sacrificando parte della varietà interna del dominio.

Questa tesi ha introdotto una serie di elementi originali nello studio delle GAN applicate alla generazione di immagini sintetiche di melanomi, distinguendosi rispetto ai lavori precedenti per l'approccio metodologico e la profondità dell'analisi. I principali contributi innovativi sono riconducibili a tre direttrici fondamentali:

- Sviluppo di un dataset clinicamente mirato: è stato creato un dataset originale, costruito e filtrato sulla base di studi scientifici specifici riguardanti le sottocategorie di nevi benigni con potenziale evolutivo verso il melanoma. Questo ha permesso di ottenere un insieme di dati non solo ricco, ma anche specificamente adatto al task generativo, garantendo una maggiore rilevanza clinica e semantica delle immagini sintetiche prodotte.

- Progettazione di un sistema di valutazione avanzato: oltre alle metriche standard comunemente utilizzate (es. FID, KID), è stato ideato un sistema di analisi che integra metriche adattate al dominio dermoscopico, capaci di cogliere aspetti come la qualità percepita e la fedeltà semantica delle immagini generate. L'integrazione di classificatori riaddestrati nel processo valutativo ha inoltre permesso di trarre conclusioni complementari e coerenti, evidenziando in modo diretto il grado di trasferimento delle feature cliniche rilevanti nelle immagini sintetiche.

- Confronto sistematico tra metodi di raffinamento del training set: è stato condotto un confronto approfondito tra diverse strategie di selezione dei dati di input per la rete generativa. In particolare, il metodo cluster-based rappresentativo, basato sui centroidi dei cluster, ha mostrato prestazioni superiori rispetto al campionamento random comunemente adottato in letteratura. Questo approccio, non documentato in precedenti studi sulle GAN per lesioni cutanee, ha permesso di generare immagini estremamente realistiche e morfologicamente fedeli, con una riproduzione efficace delle caratteristiche diagnostiche salienti.

Sviluppi futuri: Il lavoro apre a diverse direzioni di approfondimento, sia tecniche sia applicative. Un primo ambito riguarda l'impiego del dataset sintetico, che potrebbe evolvere in un vero e proprio benchmark pubblico per analizzare in modo più sistematico la progressione benigno–melanoma. Inoltre, l'integrazione con dataset dinamici reali permetterebbe di studiare la trasformazione delle lesioni in modo più fedele.

Parallelamente, le immagini generate possono essere utilizzate per strategie di data augmentation controllata, contribuendo a riequilibrare dataset sbilanciati.
Ulteriori miglioramenti riguardano la qualità visiva attraverso l'aumento della risoluzione, la riduzione degli artefatti e la normalizzazione cromatica. In questa direzione, sarà utile esplorare modelli generativi più moderni, come StyleGAN e Diffusion Models, nonché tecniche di campionamento più efficaci.
Anche le metriche di valutazione potranno essere affinate, ad esempio con un riaddestramento ancora più mirato dei feature extractor o introducendo nuove misure in grado di catturare meglio gli aspetti clinici. Un confronto diretto con dermatologi potrà infine fornire una valutazione percettiva cruciale per validare l'utilità del modello.
In un'ottica più ampia, l'intero framework potrà essere testato su nuovi dataset per valutarne la capacità di generalizzazione e potrà contribuire sia allo sviluppo di strumenti diagnostici più robusti sia al supporto clinico e alla prevenzione.

Appendice A

AizoOn Technology Consulting

AizoOn Technology Consulting è una società indipendente di tecnologia e consulenza, leader del progetto IPeR e all'interno della quale questa tesi è stata sviluppata. Ha iniziato la sua attività nel 2005 a Torino, con l'obiettivo di apportare innovazione nei settori tecnologici e industriali attraverso soluzioni digitali avanzate. La società ha esteso rapidamente la propria presenza sul territorio italiano, con sedi a Torino, Cuneo, Genova, Milano, Bologna, Roma, Bari e Catania, e si prepara ad aprire una nuova sede in Umbria. L'espansione internazionale è iniziata nel 2010 con l'apertura di filiali in Australia e negli Stati Uniti, proseguendo nel 2020 con l'ingresso in Svizzera e l'ampliamento in Europa (Francia, Spagna, Belgio) e Sud America.

La mission di aizoOn si articola su tre pilastri principali:

- **Innovation Streams**, che include cinque aree di interesse multi-settoriale in cui diversi attori condividono obiettivi comuni, attraverso ecosistemi digitali. Tra queste aree, spiccano *Digital Health*, *Digital Food*, *Environment*, *Smart Communities* e *Space4humans*;
- **Market Divisions**, che coprono una varietà di settori, tra cui *Finance*, *Government*, *Aerospace*, *Defence*, *Transportation*, *Consumer Goods & Services*, *Industrial Goods & Communication*, *Energy*, e *Health & Lifescience*;
- **Technology Divisions**, che comprendono la *Digital Engineering & Innovation Division* e la *Cyber Security Division.*

AizoOn ha scelto di concentrarsi in particolare su settori innovativi come la *Cyber Security*, l' *Intelligenza Artificiale* (AI), l'*Internet of Things* (IoT), il *Cloud Computing* e l'automazione, a cui si aggiungono nuove aree emergenti come la piattaforma

di sviluppo *Low-Code* e l'automazione dei processi. Il continuo investimento nella formazione del personale e nella collaborazione con università e centri di ricerca è fondamentale per aizoOn, che collabora attivamente con istituti come l'Università di Torino, particolarmente nel campo dell'oncologia di precisione e delle Life Sciences.

Nell'ambito del progetto IPeR, ha il compito di sviluppare algoritmi avanzati di *Machine Learning* e *Artificial Intelligence* applicati alla medicina di precisione, che saranno cruciali per il successo di IPeR, soprattutto nella creazione dei *Digital Twin* di malattia e di paziente. In particolare, la Divisione *DIGei* (Digital Engineering and Innovation), composta da un team di oltre sessanta Data Scientist e Data Engineers, è specializzata nell'analisi e nell'estrazione di valore dei dati.

Grazie ad aizoOn ho avuto l'opportunità di avvicinarmi per la prima volta a un contesto lavorativo verso cui ambisco e di sperimentarmi direttamente nell'utilizzo di algoritmi generativi, mettendomi in gioco in prima persona. La presenza di persone giovani e appassionate in un ambiente così dinamico e stimolante ha rappresentato un valore aggiunto, che mi ha arricchito sia dal punto di vista umano, sia professionale, influenzando positivamente il mio approccio ai progetti, ampliando le mie conoscenze e favorendo la collaborazione con i colleghi attraverso lo scambio di idee e consigli.

Figura A.1: aizoOn Technology Consulting

Bibliografia

[1] Lennart Jütte, Sandra González-Villà, Josep Quintana, Martin Steven, Rafael Garcia e Bernhard Roth. «Integrating generative AI with ABCDE rule analysis for enhanced skin cancer diagnosis, dermatologist training and patient education». In: *Frontiers in medicine* 11 (2024), p. 1445318 (cit. alle pp. 1, 13, 47, 74, 123).

[2] Isack Farady, Elvin Nur Furqon, Chia-Chen Kuo, Yih-Kuen Jan e Chih-Yang Lin. «Pseudo Skin Image Generator (PSIG-Net): Ambiguity-free sample generation and outlier control for skin lesion classification». In: *Biomedical Signal Processing and Control* 93 (2024), p. 106112 (cit. alle pp. 1, 11, 12).

[3] Hui Wang, Qianqian Qi, Weijia Sun, Xue Li, Boxin Dong e Chunli Yao. «Classification of skin lesions with generative adversarial networks and improved MobileNetV2». In: *International Journal of Imaging Systems and Technology* 33 (apr. 2023), n/a–n/a. DOI: 10.1002/ima.22880 (cit. a p. 1).

[4] International Skin Imaging Collaboration. *ISIC 2019: Training Dataset*. Includes BCN20000, HAM10000, and MSK datasets. CC-BY-NC license. 2019. URL: https://challenge2019.isic-archive.com/ (cit. alle pp. 2, 21, 25).

[5] International Skin Imaging Collaboration. *SIIM-ISIC 2020 Challenge Dataset*. Creative Commons Attribution–Non Commercial 4.0 International License. Images provided by Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, The University of Queensland, University of Athens Medical School. 2020. DOI: 10.34970/2020-ds01. URL: https://doi.org/10.34970/2020-ds01 (cit. alle pp. 2, 21, 26).

[6] Paolo A. Ascierto et al. *Linee guida melanoma. Edizione 2013*. Coordinatore: Paolo A. Ascierto; Segretario scientifico: Ester Simeone. Milano, Italia: Associazione Italiana di Oncologia Medica (AIOM), 2013 (cit. alle pp. 4, 5).

[7] Robyn M. Lucas, Anthony J. McMichael, Bruce K. Armstrong e Wayne T. Smith. «Estimating the global disease burden due to ultraviolet radiation exposure». In: *International Journal of Epidemiology* 37.3 (2008), pp. 654–667. DOI: 10.1093/ije/dyn017 (cit. a p. 4).

[8] International Agency for Research on Cancer (IARC). *Melanoma Awareness Month 2022*. https://www.iarc.who.int/news-events/melanoma-awareness-month-2022/. Published: 2 May 2022. Accessed: 2025-09-29. 2022 (cit. a p. 4).

[9] Shaowei Wu, Jiali Han, Francine Laden e Abrar A. Qureshi. «Long-term ultraviolet flux, other potential risk factors, and skin cancer risk: a cohort study». In: *Cancer Epidemiology, Biomarkers & Prevention* 23.6 (2014), pp. 1080–1089. DOI: 10.1158/1055-9965.EPI-13-0821 (cit. a p. 4).

[10] World Health Organization. *Cancer statistics, 2020*. https://www.who.int/publications/i/item/9789240000000. Accessed: 2025-11-13. 2020 (cit. a p. 4).

[11] C. M. Olsen, L. F. Wilson, A. C. Green, N. Biswas, J. Loyalka e D. C. Whiteman. «How many melanomas might be prevented if more people applied sunscreen regularly?» In: *British Journal of Dermatology* 178.1 (2018). Epub 2017 Dec 14, pp. 140–147. DOI: 10.1111/bjd.16079 (cit. a p. 4).

[12] International Agency for Research on Cancer (IARC). *Global burden of cutaneous melanoma in 2020 and projections to 2040*. https://www.iarc.who.int/infographics/global-burden-of-cutaneous-melanoma-in-2020-and-projections-to-2040/. Accessed: 2025-09-29. 2022 (cit. a p. 4).

[13] Jinghui Guan, Rajan Gupta e Fabian V. Filipp. «Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma». In: *Scientific Reports* 5 (2015), p. 7857. DOI: 10.1038/srep07857 (cit. a p. 5).

[14] Darrell S Rigel, Robert J Friedman, Alfred W Kopf e David Polsky. «ABCDE—an evolving concept in the early detection of melanoma». In: *Archives of dermatology* 141.8 (2005), pp. 1032–1034 (cit. a p. 5).

[15] American Cancer Society. *Tests For Melanoma Skin Cancer | Melanoma Diagnosis*. https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/how-diagnosed.html. Accessed: 2025-09-29. 2025 (cit. alle pp. 5, 6).

[16] S. R. Grant, T. W. Andrew, E. V. Alvarez, W. J. Huss e G. Paragh. «Diagnostic and Prognostic Deep Learning Applications for Histological Assessment of Cutaneous Melanoma». In: *Cancers* 14.24 (2022), p. 6231. DOI: 10.3390/cancers14246231 (cit. a p. 6).

[17] Rui Wang, Xiaofei Chen, Xiangyang Wang, Haiquan Wang, Chunhua Qian, Liucheng Yao e Kecheng Zhang. «A novel approach for melanoma detection utilizing GAN synthesis and vision transformer». In: *Computers in Biology and Medicine* 176 (2024), p. 108572 (cit. alle pp. 6, 11, 12).

[18] Hoda Naseri e Ali A Safaei. «Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: a systematic literature review». In: *BMC cancer* 25.1 (2025), p. 75 (cit. alle pp. 6, 11).

[19] Columbus3C Medical Center. *Regola ABCDE del melanoma*. Immagine tratta dal sito Columbus3C, consultato il 24 ottobre 2025. 2023. URL: https://www.columbus3c.com/blog/approfondimenti/abcde-melanoma-milano/ (cit. a p. 6).

[20] Rony Shreberk-Hassidim, Stephen M. Ostrowski e David E. Fisher. «The Complex Interplay between Nevi and Melanoma: Risk Factors and Precursors». In: *International Journal of Molecular Sciences* 24.3541 (2023). DOI: 10.3390/ijms24043541. URL: https://www.mdpi.com/1422-0067/24/4/3541 (cit. a p. 7).

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville e Yoshua Bengio. «Generative Adversarial Nets». In: *Advances in Neural Information Processing Systems*. A cura di Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence e K.Q. Weinberger. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf (cit. alle pp. 11, 15, 16, 19, 51).

[22] J. Y. Zhu, T. Park, P. Isola e A. A. Efros. «Unpaired image-to-image translation using cycle-consistent adversarial networks». In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE. 2017, pp. 2223–2232 (cit. alle pp. 11, 33–35, 38, 46, 47, 49, 50, 53, 123).

[23] Haroon Rashid, M Asjid Tanveer e Hassan Aqeel Khan. «Skin lesion classification using GAN based data augmentation». In: *2019 41St annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2019, pp. 916–919 (cit. a p. 11).

[24] Zhiwei Qin, Zhao Liu, Ping Zhu e Yongbo Xue. «A GAN-based image synthesis method for skin lesion classification». In: *Computer methods and programs in biomedicine* 195 (2020), p. 105568 (cit. alle pp. 11, 12, 66).

[25] Mudassir Saeed, Asma Naseer, Hassan Masood, Shafiq Ur Rehman e Volker Gruhn. «The power of generative ai to augment for enhanced skin cancer classification: A deep learning approach». In: *IEEE Access* 11 (2023), pp. 130330–130344 (cit. alle pp. 11, 12).

[26] Massimo Salvi, Francesco Branciforti, Federica Veronese, Elisa Zavattaro, Vanessa Tarantino, Paola Savoia e Kristen M Meiburger. «DermoCC-GAN: A new approach for standardizing dermatological images using generative adversarial networks». In: *Computer methods and programs in biomedicine* 225 (2022), p. 107040 (cit. alle pp. 11, 12).

[27] Massimo Salvi, Francesco Branciforti, Filippo Molinari e Kristen M Meiburger. «Generative models for color normalization in digital pathology and dermatology: Advancing the learning paradigm». In: *Expert Systems with Applications* 245 (2024), p. 123105 (cit. alle pp. 11, 12).

[28] Sandra Carrasco Limeros, Sylwia Majchrowska, Mohamad Khir Zoubi, Anna Rosén, Juulia Suvilehto, Lisa Sjöblom e Magnus Kjellberg. «Assessing gan-based generative modeling on skin lesions images». In: *Machine intelligence and digital interaction conference*. Springer Nature Switzerland Cham. 2022, pp. 93–102 (cit. alle pp. 11, 12, 63).

[29] Alessio Luschi, Linda Tognetti, Alessandra Cartocci, Gabriele Cevenini, Pietro Rubegni e Ernesto Iadanza. «Advancing synthetic data for dermatology: GAN comparison with multi-metric and expert validation approach». In: *Health and Technology* 15.3 (2025), pp. 553–562 (cit. alle pp. 11, 12, 62, 66, 81, 90, 93, 95, 115).

[30] Luwei Sun, Dongrui Shen e Han Feng. *Theoretical Insights into CycleGAN: Analyzing Approximation and Estimation Errors in Unpaired Data Generation*. 2025. arXiv: 2407.11678 [cs.LG]. URL: https://arxiv.org/abs/2407.11678 (cit. alle pp. 17, 33, 34, 36, 38).

[31] Poonam Chaudhari, Himanshu Agrawal e Ketan Kotecha. «Data augmentation using MG-GAN for improved cancer classification on gene expression data». In: *Soft Computing* 24 (ago. 2020), pp. 1–11. DOI: 10.1007/s00500-019-04602-2 (cit. a p. 19).

[32] International Skin Imaging Collaboration (ISIC). *ISIC Archive*. https://www.isic-archive.com. Accessed: 2025-08-23 (cit. a p. 22).

[33] Madiha Hameed, Aneela Zameer e Muhammad Asif Zahoor Raja. «A Comprehensive Systematic Review: Advancements in Skin Cancer Classification and Segmentation Using the ISIC Dataset». In: *Computer Modeling in Engineering & Sciences* 140 (2024). DOI: 10.32604/cmes.2024.050124 (cit. alle pp. 22–24, 32).

[34] Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska e Moi Hoon Yap. «Analysis of the ISIC image datasets: Usage, benchmarks and recommendations». In: *Medical Image Analysis* 75 (2022), p. 102305. ISSN: 1361-8415. DOI: 10.1016/j.media.2021.102305. URL: https://www.sciencedirect.com/science/article/pii/S1361841521003509 (cit. alle pp. 23, 25, 26, 31, 74).

[35] Keiron O'Shea e Ryan Nash. «An Introduction to Convolutional Neural Networks». In: *arXiv preprint arXiv:1511.08458* (2015) (cit. alle pp. 42, 43).

[36] Arohan Ajit, Koustav Acharya e Abhishek Samanta. «A review of convolutional neural networks». In: *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE. 2020, pp. 1–5 (cit. a p. 42).

[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren e Jian Sun. «Deep residual learning for image recognition». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. alle pp. 43, 59, 60).

[38] GeeksforGeeks. *Residual Networks (ResNet) — Deep Learning*. `https://www.geeksforgeeks.org/deep-learning/residual-networks-resnet-deep-learning/`. Accessed: 2025-09-30. 2023 (cit. a p. 43).

[39] Dmitry Ulyanov, Andrea Vedaldi e Victor Lempitsky. «Instance Normalization: The Missing Ingredient for Fast Stylization». In: *arXiv preprint arXiv:1607.08022* (2017) (cit. a p. 44).

[40] Christian Schneppat. *Instance Normalization*. `https://schneppat.com/instance-normalization_in.html`. Accessed: 2025-09-30. 2023 (cit. alle pp. 44, 45).

[41] Vinod Nair e Geoffrey E Hinton. «Rectified linear units improve restricted boltzmann machines». In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 2010, pp. 807–814 (cit. a p. 45).

[42] Xavier Glorot, Antoine Bordes e Yoshua Bengio. «Deep sparse rectifier neural networks». In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. 2011, pp. 315–323 (cit. a p. 45).

[43] Andrew L Maas, Awni Y Hannun e Andrew Y Ng. «Rectifier nonlinearities improve neural network acoustic models». In: *Proceedings of the 30th International Conference on Machine Learning*. 2013 (cit. a p. 46).

[44] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou e Alexei A. Efros. «Image-to-image translation with conditional adversarial networks». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134 (cit. a p. 46).

[45] Alec Radford, Luke Metz e Soumith Chintala. «Unsupervised representation learning with deep convolutional generative adversarial networks». In: 2015 (cit. a p. 46).

[46] Chilldenaya. *CycleGAN Introduction & PyTorch Implementation*. `https://medium.com/@chilldenaya/cyclegan-introduction-pytorch-implementation-5b53913741ca`. Accessed: 2025-11-12. 2020 (cit. alle pp. 49, 50).

[47] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang e Stephen Paul Smolley. «Least squares generative adversarial networks». In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2794–2802 (cit. alle pp. 50, 51, 55).

[48] Vibolroth Sambath, Nicolas Viltard, Laurent Barthès, Audrey Martini e Cécile Mallet. «Unsupervised domain adaptation for global precipitation measurement satellite constellation using cycle generative adversarial nets». In: *Environmental Data Science* 1 (2022), e24 (cit. alle pp. 54, 83).

[49] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler e Sepp Hochreiter. «Gans trained by a two time-scale update rule converge to a local nash equilibrium». In: *Advances in neural information processing systems* 30 (2017) (cit. alle pp. 57, 61, 62).

[50] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel e Arthur Gretton. «Demystifying mmd gans». In: *arXiv preprint arXiv:1801.01401* (2018) (cit. alle pp. 57, 61, 63).

[51] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen e Timo Aila. «Improved precision and recall metric for assessing generative models». In: *Advances in neural information processing systems* 32 (2019) (cit. alle pp. 57, 61, 64).

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford e Xi Chen. «Improved techniques for training gans». In: *Advances in neural information processing systems* 29 (2016) (cit. a p. 57).

[53] Yash Deo, Yan Jia, Toni Lassila, William AP Smith, Tom Lawton, Siyuan Kang, Alejandro F Frangi e Ibrahim Habli. «Metrics that matter: Evaluating image quality metrics for medical image generation». In: *arXiv preprint arXiv:2505.07175* (2025) (cit. alle pp. 57, 60, 62, 63, 65, 89).

[54] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens e Zbigniew Wojna. «Rethinking the inception architecture for computer vision». In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826 (cit. alle pp. 58, 59).

[55] Nitish Kundu. *Exploring ResNet50: An In-Depth Look at the Model Architecture and Code Implementation*. `https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f`. 2023 (cit. alle pp. 59, 60).

[56] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi e Jaejun Yoo. «Reliable fidelity and diversity metrics for generative models». In: *International conference on machine learning*. PMLR. 2020, pp. 7176–7185 (cit. alle pp. 64–66).

[57] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet e Sylvain Gelly. «Assessing generative models via precision and recall». In: *Advances in neural information processing systems* 31 (2018) (cit. a p. 64).

[58] PyTorch. *Transfer Learning Tutorial*. Accessed: 2025-10-22. 2020. URL: `https://docs.pytorch.org/tutorials/beginner/transfer_learning_tutorial.html` (cit. a p. 68).

[59] Qishen Ha, Bo Liu e Fuxu Liu. *Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge*. 2020. arXiv: `2010.05351 [cs.CV]`. URL: `https://arxiv.org/abs/2010.05351` (cit. alle pp. 70, 71, 109).

[60] Mominul Islam, Hasib Zunair e Nabeel Mohammed. «CosSIF: Cosine similarity-based image filtering to overcome low inter-class variation in synthetic medical image datasets». In: *Computers in Biology and Medicine* 172 (2024), p. 108317 (cit. a p. 75).

[61] Show-Jane Yen e Yue-Shi Lee. «Cluster-based under-sampling approaches for imbalanced data distributions». In: *Expert systems with applications* 36.3 (2009), pp. 5718–5727 (cit. a p. 75).

[62] Md Abdur Rahman, Nur Mohammad Fahad, Mohaimenul Azam Khan Raiaan, Mirjam Jonkman, Friso De Boer e Sami Azam. «Advancing skin cancer detection integrating a novel unsupervised classification and enhanced imaging techniques». In: *CAAI Transactions on Intelligence Technology* 10.2 (2025), pp. 474–493 (cit. alle pp. 75, 78).

[63] Terrance DeVries, Michal Drozdzal e Graham W Taylor. «Instance selection for gans». In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 13285–13296 (cit. alle pp. 76, 96).