POLITECNICO DI TORINO

Laurea magistrale in ingegneria biomedica

Tesi di laurea magistrale

Combination of Visual and Metadata Information for Accurate Lesion Classification using Deep Learning

Relatori

Prof.ssa Kristen MEIBURGER

Prof. Massimo SALVI

Ing. Francesco BRANCIFORTI

Candidata

Lara PISTORIO

Luglio 2024

Sommario

La pelle è l'organo più esteso del corpo umano e le patologie che possono interessarla sono molteplici. Una delle forme più comuni di cancro a livello mondiale è il cancro alla pelle e ciò mette in evidenza la necessità di possedere strumenti diagnostici atti a rilevare precocemente e con precisione la natura maligna di alcune lesioni cutanee. L'obiettivo di questo lavoro di tesi è applicare il modello multimodale CLIP (Contrastive Language-Image Pretraining) di OpenAI per automatizzare la classificazione delle immagini di lesioni cutanee in benigne e maligne.

Il punto di forza di CLIP è la sua capacità di integrare la visione artificiale alla comprensione del linguaggio naturale, consentendo al modello di combinare concetti visivi e testuali in modo sinergico. In questo studio è stato utilizzato un dataset contenente immagini dermoscopiche di lesioni cutanee e descrizioni testuali generate a partire dai metadati associati ad ogni immagine. Nello specifico, i metadati includono informazioni riguardanti il paziente come l'età e il genere e caratteristiche della lesione come la sua collocazione anatomica e la diagnosi clinica.

Durante l'allenamento del modello, ogni immagine è stata presentata insieme alla sua descrizione testuale, permettendo a CLIP di apprendere le correlazioni tra i dati visivi e le informazioni testuali. Il modello CLIP è stato allenato su un dataset di training e validato su un development set, con l'obiettivo di ottimizzare i parametri in modo tale da massimizzare l'accuratezza e la precisione della classificazione. Infine, il modello è stato testato su un dataset di test contenente immagini mai viste dal modello e le sue performance sono state valutate utilizzando metriche come l'accuratezza, la sensibilità, la specificità, la precisione e l'F1-score.

I risultati ottenuti mostrano che l'approccio basato sul modello CLIP può migliorare significativamente la rilevazione precoce del cancro alla pelle, fornendo un supporto ai dermatologi nella valutazione delle lesioni cutanee. L'utilizzo di modelli multimodali come CLIP rappresenta un'ulteriore innovazione nella pratica clinica, permettendo di giungere a diagnosi più accurate in modo tempestivo. Inoltre, questo lavoro apre nuove prospettive per nuove ricerche nel campo dell'intelligenza

artificiale applicata alla dermatologia. In conclusione, questo studio dimostra che l'utilizzo del modello CLIP per la classificazione delle lesioni cutanee potrebbe portare alla rivoluzione della metodologia secondo la quale esse vengono diagnosticate, riducendo il carico di lavoro dei dermatologi e supportandoli nella valutazione, con benefici significativi per la salute pubblica.

Ringraziamenti

Ai miei genitori, le mie colonne portanti;
A Eva, l'altra metà del mio cuore;
Ai nonni, a cui mando un bacio da qui;
A Niki e Maika, i miei momenti di spensieratezza quotidiana;
A Luca, il mio punto di riferimento e la mia forza.

Indice

Elenco delle tabelle

Elenco delle figure

Acronimi

AI
: artificial intelligence

DL
: deep learning

CV
: computer vision

ViT
: vision transformer

CLIP
: contrastive language-image pretraining

ISIC
: international skin imaging collaboration

Capitolo 1

Introduzione

1.1 Contesto

L'era della digitalizzazione ha rivoluzionato innumerevoli settori, portando con sé cambiamenti profondi nel modo in cui vengono raccolti, analizzati e utilizzati i dati. Uno dei campi che ha beneficiato significativamente di questa trasformazione è quello della dermatologia, in particolare nell'analisi delle immagini di lesioni cutanee, aprendo così a nuove frontiere nella diagnosi, nella ricerca e nella gestione delle malattie cutanee.

I metodi tradizionali per l'analisi delle lesioni cutanee si basano principalmente su perizie dermatologiche ed esami istopatologici, che possono richiedere molto tempo e risultare soggetti a variabilità nell'accuratezza diagnostica.

Negli ultimi anni, i progressi nel Deep Learning e nella Computer Vision hanno portato a breakthrough anche del dominio delle immagini mediche, con l'obiettivo di di poter offrire strumenti diagnostici più rapidi, affidabili e accessibili, specialmente nelle regioni con accesso limitato a dermatologi specializzati.

Le prime ricerche nel settore si sono concentrate principalmente sull'elaborazione delle immagini mediche. Questo approccio ha consentito lo sviluppo di sistemi capaci di riconoscere pattern e anomalie visive con crescente accuratezza. Tuttavia, presentava limitazioni in termini di contestualizzazione clinica, non considerando informazioni cruciali al di fuori dell'immagine stessa.

Recentemente, la ricerca si è orientata verso l'integrazione di informazioni testuali, quali schede cliniche e metadati, nel processo di analisi attraverso modelli di Deep Learning che vengono perciò detti "multimodali".[1]

Questi approcci permetterebbero di fornire ai sistemi di Deep Learning un contesto più ricco e clinicamente rilevante, garantendo una migliorare comprensione delle correlazioni tra sintomi visivi e dati clinici.

Uno dei contributi più significativi è rappresentato da CLIP (Contrastive Language-Image Pretraining) di OpenAI, un modello multimodale pre-addestrato che unisce l'analisi delle immagini alla comprensione del testo. [2] CLIP si distingue per le sue proprietà emergenti, che ne consentono l'applicazione su una vasta gamma di task di Machine Learning, inclusa la classificazione.

1.2 Struttura della tesi

La presente ricerca si propone di esplorare e validare l'efficacia del fine tuning di CLIP per la classificazione delle lesioni cutanee.
La trattazione sarà distribuita nei vari capitoli come segue:

- **Capitolo 2**: viene presentato lo stato dell'arte nell'ambito della Computer Vision applicata al contesto biomedicale, vengono ripercorse le fondamenta teoriche dei modelli basati su architettura Transformer, inclusi i Vision Transformers e viene approfondito CLIP.
- **Capitolo 3**: viene descritto nel dettaglio l'approccio utilizzato per il fine tuning di CLIP, comprese le modalità di integrazione dei metadati e la configurazione specifica per la classificazione delle lesioni cutanee.
- **Capitolo 4**: vengono presentati i risultati delle simulazioni effettuate sotto diverse configurazioni del modello, analizzando le prestazioni ottenute in base all'efficacia dei metadati utilizzati.
- **Capitolo 5**: vengono discussi i risultati della ricerca, analizzando le implicazioni e le possibili direzioni future per la diagnosi e il trattamento delle lesioni cutanee tramite modelli multimodali.

Capitolo 2

Background

2.1 ISIC Dataset

L'International Skin Imaging Collaboration (ISIC) [3] rappresenta una risorsa di dati open access di primaria importanza nel campo della dermatologia digitale. Il database ISIC si distingue per la sua vasta collezione di immagini dermatologiche ad alta risoluzione, accompagnate da metadati dettagliati che includono diagnosi istologiche, localizzazione anatomica delle lesioni e informazioni demografiche dei pazienti. Secondo quanto riportato dal loro sito, al mese di luglio 2024 il database contiene oltre 81.000 immagini di lesioni cutanee disponibili pubblicamente, di cui circa l'11% mostra casi di melanoma e altri tumori cutanei maligni.

L'archivio ISIC mantiene la sua rilevanza attraverso aggiornamenti regolari, garantendo che il dataset rimanga una risorsa attuale e preziosa per ricercatori e clinici. L'organizzazione promuove inoltre l'innovazione nel settore attraverso challenge annuali, come l'ISIC Challenge, che nel 2020 ha visto la partecipazione di oltre 3.000 partecipanti da tutto il mondo.

Il dataset ISIC non si limita a facilitare l'addestramento e la valutazione di modelli di classificazione, ma riflette anche le complessità della pratica clinica reale. La diversità nelle tecniche di acquisizione delle immagini e nelle caratteristiche dei pazienti rappresentati nel dataset - che include soggetti di varie etnie e fasce d'età - pone sfide significative nello sviluppo di algoritmi diagnostici robusti.

2.2 Sfide nello studio delle lesioni cutanee

La classificazione delle lesioni cutanee mediante sistemi automatizzati deve affrontare diverse sfide, ampiamente documentate in letteratura. Un problema primario è l'elevato grado di eterogeneità visiva all'interno delle categorie di lesioni. I melanomi, ad esempio, presentano un'ampia gamma di morfologie, colori e irregolarità dei bordi. Inoltre, le lesioni benigne possono spesso imitare quelle maligne, portando ad un alto tasso di falsi positivi nei sistemi di classificazione meno sofisticati. [4]

Un'altra sfida significativa è la presenza di artefatti nelle immagini dermoscopiche, come peli, bolle e ombre, che possono oscurare le caratteristiche della lesione o essere interpretati erroneamente come parte della lesione, confondendo così il processo di classificazione. La diversa qualità e risoluzione delle immagini, a seconda del dispositivo e della tecnica di acquisizione, complica ulteriormente il compito. [5]

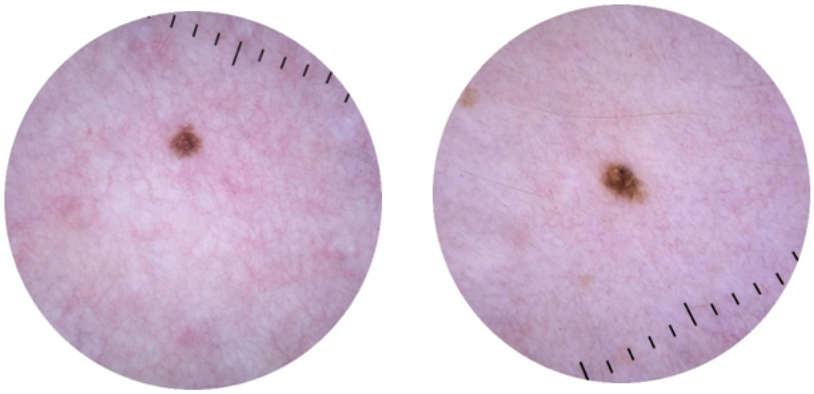

Figura 2.1: Esempio di come lesioni benigne (prima immagine) e lesioni maligne (seconda immagine) siano simili

Si sottolinea anche il problema dello sbilanciamento delle classi, un problema comune nei set di dati medici in cui alcune condizioni sono naturalmente più rare di altre. Questo squilibrio può alterare le prestazioni di un modello, rendendolo più abile nel riconoscere le condizioni prevalenti, mentre potrebbe mostrare più difficoltà con le lesioni rare, ma spesso più pericolose. [6]

Il contributo che potrebbero apportare i metadati, come l'età del paziente, la posizione anatomica della lesione e l'anamnesi del cancro della pelle, andrebbe a beneficio delle prestazioni di classificazione. In generale, l'utilizzo di metadati ed

annotazioni può diventare un ottimo discriminante per distinguere ed identificare lesioni maligne.

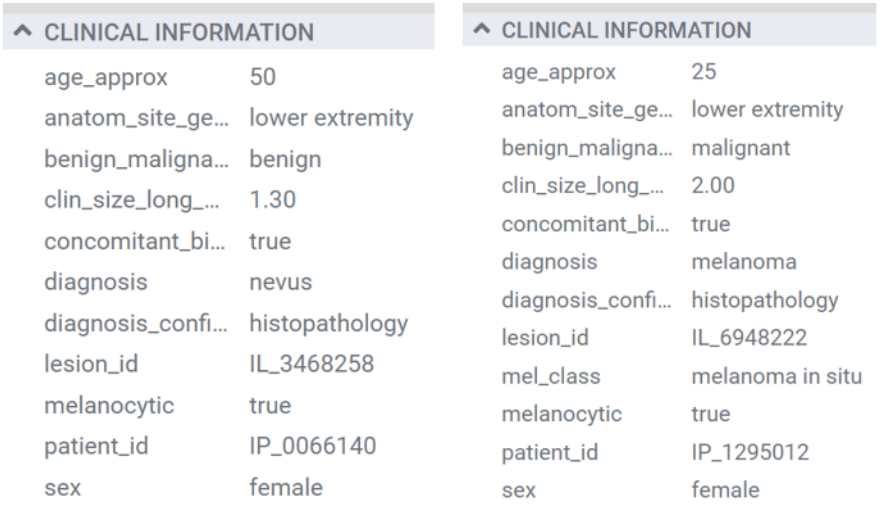

Figura 2.2: Metadati associati alla lesione benigna (prima immagine) e alla lesione maligna (seconda immagine) Fonte: ISIC[3]

2.3 Metodi tradizionali

Quando si tratta di classificare le lesioni cutanee, ci sono vari metodi tradizionali che vengono utilizzati. Vediamone alcuni:

1. **Regola ABCDE [7]**: questo metodo valuta le lesioni cutanee basandosi su cinque caratteristiche:
   (a) Asimmetria,
   (b) Bordi irregolari,
   (c) Colore non uniforme,
   (d) Diametro superiore ai 6 mm ed
   (e) Evoluzione nel tempo.

   È un modo molto diffuso per identificare potenziali melanomi [8]

2. **Metodo di Pattern Analysis**: con questo metodo si analizzano i pattern globali e locali delle lesioni dermoscopiche. L'idea è quella di identificare specifiche strutture che possono essere associate a lesioni benigne o maligne.

3. **7-point checklist[9, 10]**: questo metodo valuta sette criteri dermoscopici specifici, assegnando punteggi per determinare la probabilità che la lesione sia un melanoma. È una checklist che aiuta a identificare le lesioni sospette in modo sistematico.

4. **Metodo di Menzies[11]**: questo approccio dettagliato e approfondito si basa sia su criteri negativi (assenza di determinate caratteristiche) che su criteri positivi (presenza di specifiche caratteristiche) per identificare un potenziale melanoma.

5. **Algoritmo CASH (Color, Architecture, Symmetry, Homogeneity)[12]**: questo algoritmo valuta quattro parametri principali delle lesioni cutanee: Colore, Architettura, Simmetria e Omogeneità. È un metodo che cerca di semplificare il processo di diagnosi mantenendo un buon livello di accuratezza.

2.4 Deep Learning

I recenti progressi nella classificazione delle lesioni cutanee hanno evidenziato che il contributo delle tecniche di Machine Learning (ML) e di Deep Learning (DL) può essere promettente, in particolar modo per l'individuazione di lesioni cutanee maligne. Tra queste tecniche, le reti neurali convoluzionali (CNN) hanno dimostrato una notevole efficacia, raggiungendo livelli di accuratezza paragonabili a quelli dei dermatologi [13, 14]. Varie architetture di CNN, come la VGG, l'Inception, la ResNet e la DenseNet, sono state ampiamente analizzate ed utilizzate per l'estrazione delle feature di un'immagine [14].
Queste architetture sono spesso abbinate anche a classificatori tradizionali come le Support Vector Machine (SVM), Random Forests e k-Nearest Neighbors (k-NN) per incrementare ulteriormente le prestazioni della rete [14, 15]

Il Transfer Learning, che sfrutta modelli pre-allenati su grandi set di dati, ha dimostrato di essere particolarmente efficace nel migliorare l'accuratezza della classificazione nel rilevamento delle lesioni cutanee maligne [14]. Questo approccio risulta essere vantaggioso a causa della limitata disponibilità di immagini mediche annotate, in quanto consente ai modelli di generalizzare meglio da altri insiemi di dati, trasferendo le conoscenze da domini diversi.

Sebbene i metodi di Deep Learning, in particolare le CNN, dominino il campo, vengono impiegate anche tecniche di ML tradizionali e approcci combinati ML/DL. I metodi ibridi combinano i punti di forza di entrambe le tecniche, spesso portando a un miglioramento delle prestazioni e della robustezza
. Ad esempio, i metodi Ensemble che combinano più modelli DL sono stati utilizzati per ottenere una maggiore accuratezza e affidabilità nella classificazione del cancro della pelle [16].

2.5 CLIP

Per introdurre l'architettura di CLIP, è importante prima comprendere i Transformers, che costituiscono la base fondamentale di molti modelli avanzati di intelligenza artificiale, inclusi quelli utilizzati in CLIP.

2.5.1 Transformers

I Transformer rappresentano una pietra miliare nel campo dell'intelligenza artificiale, essendo un'architettura neurale trasformativa che ha rivoluzionato il modo in cui i modelli possono elaborare e comprendere il linguaggio naturale e altri tipi di dati sequenziali. Originariamente introdotti nel 2017 da Vaswani et al. nel paper *Attention is all you need* [17], i Transformer hanno superato le limitazioni dei modelli precedenti grazie alla loro capacità di modellare relazioni complesse e a lungo raggio tra token all'interno di sequenze.

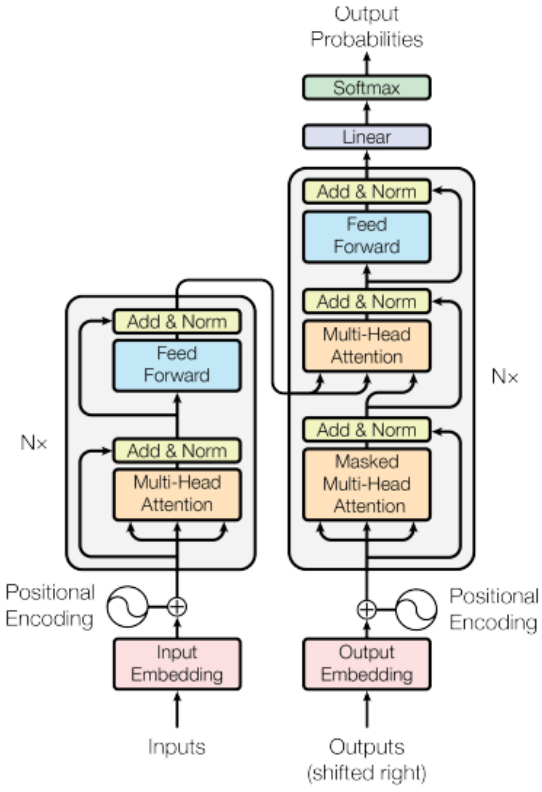

Figura 2.3: Modello architettura rete Transformer

Al cuore dei Transformer ci sono meccanismi di attenzione multi-head, che permettono al modello di focalizzare l'attenzione su parti specifiche della sequenza durante il processo di elaborazione. Questo approccio consente una parallelizzazione efficace, rendendo i Transformer scalabili anche su dataset di grandi dimensioni. Inoltre, l'uso di strati di self-attention rende possibile l'apprendimento di relazioni semantiche e contestuali tra token, migliorando le prestazioni su compiti di comprensione del linguaggio naturale come traduzione, generazione di testo e classificazione.

2.5.2 Vision Transformers

I vision transformers sono una versione adattata dei transformer originali, ottimizzata per il trattamento delle immagini anziché del testo. Questa architettura, introdotta nell'Ottobre 2020 [18], rappresenta un significativo avanzamento nel campo della visione artificiale, sostituendo le tradizionali reti neurali convoluzionali (CNN) con meccanismi di attenzione self-attention. Questi meccanismi permettono al modello di focalizzarsi su regioni specifiche dell'immagine durante l'elaborazione, in modo analogo a come i transformer trattano i token nelle sequenze di testo.

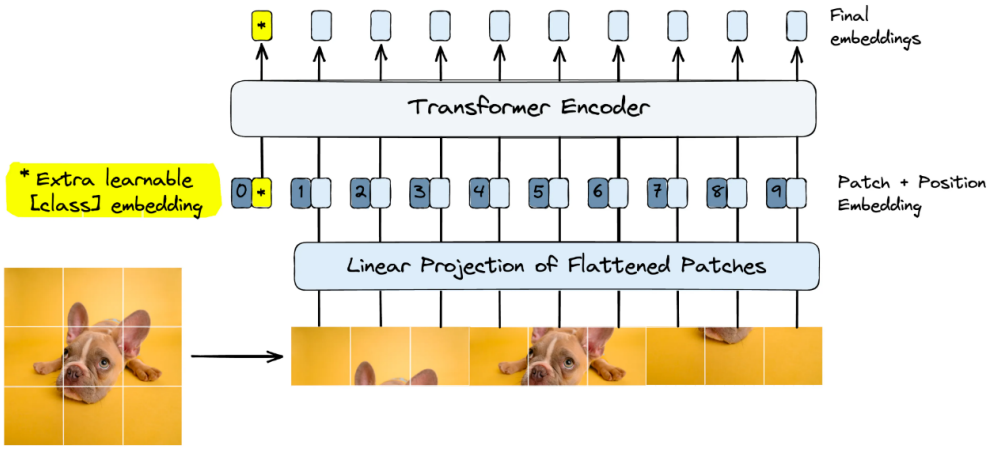

Figura 2.4: Vision Transformer [19]

Un ViT è composto da blocchi di transformer, ciascuno dei quali include strati di self-attention e reti feedforward completamente connessi. Durante l'addestramento, l'immagine viene suddivisa in patch e convertita in una sequenza di vettori, che vengono quindi elaborati dai blocchi di transformer per catturare informazioni gerarchiche e contestuali all'interno dell'immagine stessa. Questo approccio consente ai vision transformers di apprendere rappresentazioni di alto livello delle caratteristiche visive, senza la necessità di strati convoluzionali.

2.5.3 Contrastive Language Image Pretraining

Sviluppato da OpenAI nel 2021 [2], CLIP è un metodo di pre-training multimodale che apprende concetti visivi dalla supervisione del linguaggio naturale.

Il nucleo di CLIP è costituito da due componenti principali: un codificatore di immagini e un codificatore di testo. Il codificatore di immagini è tipicamente un Vision Transformer (ViT) o una rete neurale del tipo ResNet che elabora le immagini in ingresso in una serie di embeddings, o rappresentazioni di caratteristiche relative alla specifica immagine. Contemporaneamente, il codificatore di testo, che assume la forma di un modello linguistico basato su un trasformatore costituito da 12 layer, elabora il testo in ingresso, come didascalie o frasi, in un insieme separato di embeddings testuali.

L'innovazione di CLIP risiede nel come viene formalizzato la funzione obiettivo durante il traninig. In particolare, il modello viene addestrato su una gamma diversificata di immagini abbinate a descrizioni testuali, con l'obiettivo di prevedere l'abbinamento corretto tra una serie di coppie immagine-testo non corrispondenti.

Questo approccio di contrastive learning incoraggia il modello a comprendere e allineare le rappresentazioni di immagini e testi in uno spazio di embeddings condiviso. Di conseguenza, il modello apprende una vasta gamma di concetti visivi direttamente dal linguaggio naturale, senza la necessità di un'annotazione manuale esplicita delle immagini.

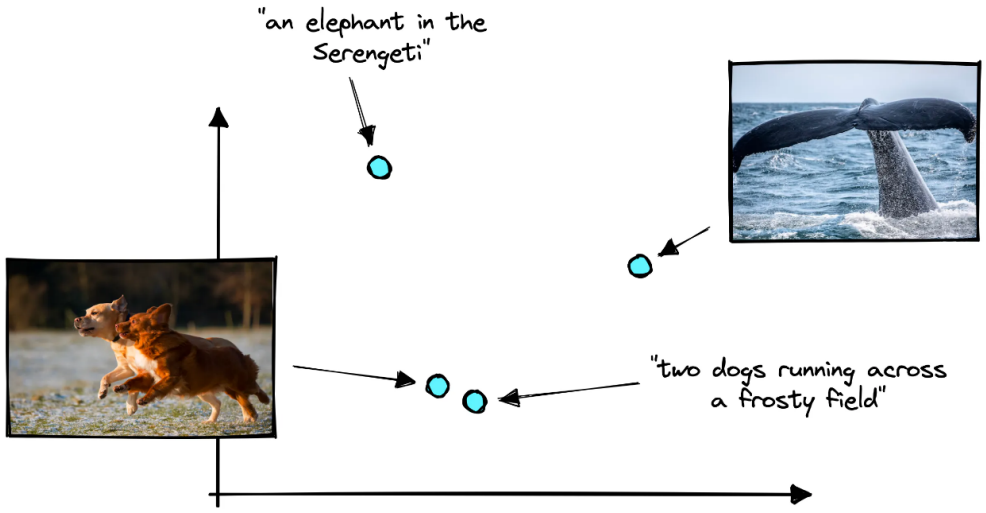

Figura 2.5: Coppie immagine-testo simili vengono codificate in uno spazio vettoriale simile [20]

2.5.4 Vantaggi di CLIP

Le implicazioni dell'architettura di CLIP per la classificazione delle immagini sono notevoli. I modelli tradizionali di classificazione delle immagini richiedono grandi insiemi di dati etichettati, in cui ogni immagine è annotata con una specifica etichetta di classe. CLIP, invece, può essere addestrato su immagini accompagnate da descrizioni testuali, spesso più facili da ottenere. Ciò consente processi di addestramento più scalabili e flessibili, in grado di sfruttare l'abbondanza di immagini descritte testualmente disponibili su internet.

Per la classificazione delle lesioni cutanee, l'architettura di CLIP è particolarmente importante a causa della natura complessa del compito. Le lesioni cutanee possono variare notevolmente nell'aspetto e la loro classificazione spesso non è semplice.
La capacità di CLIP di comprendere il contesto fornito dai metadati sotto forma di descrizioni testuali può essere determinante. Per esempio, un prompt testuale che riporta "un nevo benigno sulla pelle esposta al sole" fornisce un contesto ricco che può aiutare il modello a comprendere le sottili differenze visive che differenziano un nevo benigno da una lesione maligna.

Inoltre, la generalizzabilità delle caratteristiche apprese da CLIP significa che, una volta preaddestrato, può essere fine-tuned con un set di dati relativamente piccolo - nell'ordine delle migliaia e non di milioni di lesioni cutanee e descrizioni associate. Questo è particolarmente vantaggioso nei domini medici, dove l'acquisizione di grandi set di dati annotati è spesso impegnativa e costosa.

2.5.5 CLIP e Prompting

L'integrazione di CLIP nella classificazione delle lesioni cutanee comporta il fine tuning del modello preaddestrato con dati specifici relativi al campo di interesse. Le descrizioni, dette *prompt* nell'ambito del Natural Language Processing (NLP), possono essere generati in modo da includere metadati come la tipologia della lesione, la sua posizione a livello anatomico, l'età del paziente e la sua precedente storia clinica. In questo modo, il modello non solo apprende dalle caratteristiche visive delle lesioni, ma sfrutta anche il contesto clinico nel suo processo decisionale.

Questo approccio ha il potenziale di migliorare significativamente le prestazioni di classificazione, fornendo una comprensione più completa delle lesioni. Inoltre, il modello può spiegare le sue previsioni in modo più comprensibile per i medici, facendo riferimento alle descrizioni testuali utilizzate per effettuare le classificazioni.

2.6 Metadati nella classificazione

L'integrazione dei metadati testuali nella classificazione delle immagini è da tempo un'area di interesse, che consente ai modelli di sfruttare un contesto che va oltre le informazioni acquisibili solo a livello visivo, perché più astratte (idee e concetti). L'unione di dati visivi e testuali ha storicamente contribuito ad arricchire la classificazione delle immagini. La seguente panoramica mostra una prospettiva storica sull'uso dei metadati testuali nella classificazione delle immagini e sulla loro evoluzione nel tempo.

2.6.1 Utilizzo storico dei metadati di testo

Storicamente, l'uso dei metadati testuali nella classificazione delle immagini è stato guidato dalla necessità di fornire un contesto aggiuntivo, non immediatamente percepibile dall'immagine stessa. I primi tentativi di incorporare metadati testuali prevedevano l'annotazione manuale delle immagini con parole chiave o tag che ne descrivevano il contenuto. Queste annotazioni venivano poi utilizzate per assistere il processo di classificazione.

Nell'era dell'apprendimento automatico, prima della rivoluzione del Deep Learning, le feature visive estratte dalle immagini erano spesso combinate con caratteristiche testuali in un classificatore. Ad esempio, le Support Vector Machine (SVM [21]) sono state addestrate su feature vectors che includevano sia metadati visivi, come istogrammi di colore o caratteristiche di texture, sia metadati di testo, come tag o didascalie. Questa combinazione mirava a migliorare l'accuratezza dei classificatori fornendo loro un insieme più ricco di caratteristiche discriminanti.

2.6.2 Progressi con il Deep Learning

L'avvento del Deep Learning e lo sviluppo delle reti neurali convoluzionali (CNN) hanno segnato un cambiamento significativo nell'utilizzo dei metadati testuali nella classificazione delle immagini. Inizialmente, gli approcci di Deep Learning si sono concentrati molto sull'apprendimento di rappresentazioni direttamente dai dati dei pixel, con minore attenzione ai metadati testuali esterni. Tuttavia, quando questi modelli sono diventati più sofisticati, è emerso nuovamente il potenziale nell'incorporare i metadati testuali.

L'introduzione di modelli di Deep Learning multimodali [1], in grado di elaborare e mettere in relazione informazioni provenienti da diversi tipi di dati, come immagini e testo, ha portato a una rinascita nell'utilizzo dei metadati testuali.

2.6.3 Il Contrastive Learning

Il concetto del Constrastive Learning, incarnato dall'architettura di CLIP, ha portato il ruolo dei metadati testuali a nuovi livelli. Grazie al preaddestramento su una gamma diversificata di immagini e sulle realtive descrizioni testuali, CLIP impara a comprendere e classificare le immagini nel contesto fornito dal testo. Questo approccio si differenzia in modo significativo dai metodi precedenti, in quanto non si basa esclusivamente su annotazioni manuali o combinazioni di caratteristiche.

L'uso dei metadati testuali in CLIP e in modelli simili va oltre la semplice corrispondenza di parole chiave o l'aumento del vettore di caratteristiche. Si tratta invece di comprendere la relazione semantica tra testo e immagine, consentendo al modello di generalizzare dai metadati a nuove immagini mai viste. Questo ha il potenziale per trasformare la classificazione delle immagini, soprattutto in campi specializzati come l'imaging medico, dove il contesto fornito dai metadati è spesso fondamentale per una diagnosi accurata.

2.6.4 L'impatto sulla classificazione delle immagini mediche

Nella classificazione delle immagini mediche, i metadati testuali contengono spesso informazioni preziose sull'anamnesi del paziente, sul tipo di imaging medico utilizzato e sulle annotazioni o diagnosi degli esperti. L'incorporazione di questi metadati consente di creare modelli più sfumati in grado di tenere conto di fattori quali i cambiamenti dei tessuti legati all'età, le variazioni nelle modalità di imaging e le sottili distinzioni tra condizioni che possono sembrare visivamente simili.

Il ruolo storico dei metadati testuali nella classificazione delle immagini dimostra una traiettoria che va dall'annotazione manuale verso una comprensione più sofisticata e automatizzata delle relazioni visive-testuali. Questa evoluzione ha avuto il culmine in architetture come CLIP, che rappresentano l'avanguardia dell'apprendimento multimodale. Con il progredire del settore, l'integrazione dei metadati testuali è destinata a diventare sempre più parte integrante della classificazione delle immagini, offrendo la promessa di modelli in grado di comprendere le immagini con una profondità e un contesto che rispecchiano la percezione umana.

Capitolo 3

Metodologia

3.1 Raccolta dei dati

Il punto di partenza di un qualsiasi progetto di apprendimento automatico è un set di dati che sia completo e rappresentativo del problema e la qualità e la quantità dei dati su cui viene addestrato. Nel contesto di questo studio, la raccolta dei dati ha comportato l'acquisizione di immagini e metadati dal dataset dell'Internation Skin Imaging Collaboration (ISIC), disponibili pubblicamente e ampiamente utilizzati dalla comunità di ricerca dermatologica per lo sviluppo di sistemi diagnostici automatizzati. Questa sezione illustra la metodologia utilizzata per raccogliere e preparare i dati per la successiva fase di analisi.

3.1.1 Acquisizione da ISIC

ISIC Archive [3], una risorsa di dominio pubblico, è un ricco archivio di immagini di lesioni cutanee, che è stato fondamentale per fare progredire la ricerca sull'apprendimento automatico nel campo della dermatologia. Il processo di raccolta dei dati da ISIC ha comportato diverse fasi:

1. **Accesso all'archivio**: l'archivio ISIC è accessibile attraverso una piattaforma online, che fornisce agli utenti un'interfaccia per cercare, visualizzare e scaricare immagini dermoscopiche di alta qualità insieme ai corrispondenti metadati clinici.

2. **Selezione delle immagini**: le immagini sono state selezionate in base a criteri predefiniti rilevanti per lo studio, come la metodologia di acquisizione dell'immagine, nel nostro caso specifico sono state prese in considerazione soltanto immagini ottenute tramite dermatoscopia. Questo processo di selezione ha garantito che il set di dati fosse rilevante e favorevole all'addestramento di un classificatore accurato.

3. **Raccolta dei metadati**: oltre alle immagini, sono stati considerati anche i metadati associati. Questi metadati includono tipicamente i dati demografici del paziente, la posizione anatomica e la presenza di melanociti nella lesione e le informazioni diagnostiche. I metadati forniscono un contesto cruciale che aiuta il compito di classificazione, in particolare quando si utilizzano modelli come CLIP che sfrutteranno le descrizioni testuali in combinazione con le informazioni estratte dalle immagini.

4. **Pre-elaborazione dei dati**: prima di essere forniti in input al modello, le immagini e i metadati raccolti sono stati sottoposti ad una fase di pre-elaborazione per garantire la compatibilità con i requisiti richiesti in input da CLIP. Questa fase ha comportato operazioni quali il ridimensionamento delle immagini, il successivo center crop, la normalizzazione dei valori dei pixel per ogni canale colore e la codifica dei metadati in un formato adatto all'addestramento del modello.

5. **Suddivisione del dataset**: i dati raccolti sono stati poi suddivisi in un training set, un development set e un test set, una fase cruciale che garantisce che il modello possa essere addestrato e valutato su set diversi di dati. È stato utilizzato un campionamento randomico, in dettaglio il 60% dei dati è stato inserito nel training set, il 20% nel development set e il restante 20% nel test set.

6. **Data Augmentation**: per risolvere il problema dello sbilanciamento delle classi dovuto alla presenza di un numero preponderante di immagini di lesioni benigne e migliorare le capacità di generalizzazione del modello, sono state applicate alle immagini della classe meno rappresentata delle tecniche di data augmentation. Queste tecniche includevano rotazioni e flip orizzontali e verticali dell'immagine, in modo tale da ottenere un maggior numero di esempi per la classe maligna.

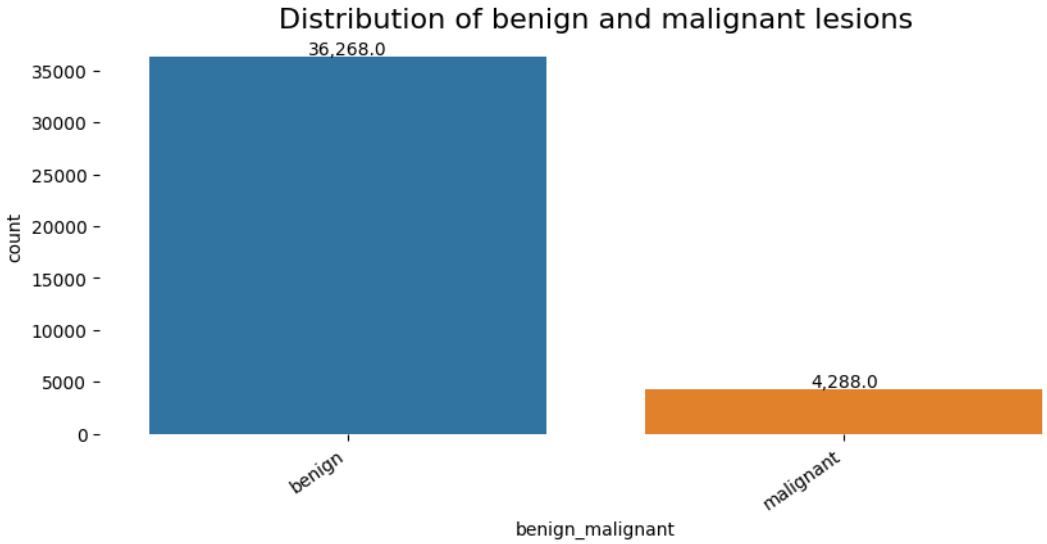

Figura 3.1: Sbilanciamento delle classi per il training set

3.1.2 Criteri di selezione

Ai fini di questo studio, è stato selezionato un sottoinsieme specifico del set di dati ISIC in base ai seguenti criteri:

1. **Tipo di immagine**: sono state considerate solo le immagini di tipo dermoscopico. L'imaging dermoscopico, una tecnica non invasiva, offre una migliore visualizzazione delle caratteristiche superficiali e subsuperficiali delle lesioni cutanee. Questa scelta garantisce una maggiore qualità e coerenza dei dati visivi, fondamentale per l'accuratezza dei modelli di classificazione basati sulle immagini.

2. **Disponibilità delle etichette di classe**: era indispendabile che ogni immagine del set di dati avesse un'etichetta chiaramente definita che indicasse la benignità o la malignità della lesione cutanea. Le immagini prive di questa informazione sono state escluse per mantenere l'integrità e l'affidabilità dei dati di addestramento.

3. **Gestione delle etichette indeterminate**: i casi etichettati come "indeterminati/benigni" sono stati ricategorizzati come "benigni", mentre quelli contrassegnati come "indeterminati/maligni" sono stati rietichettati come "maligni". Questa decisione è stata presa per mantenere un sistema di classificazione binario, essenziale per l'esecuzione del task di classificazione.

3.1.3 Inclusione dei metadati

Insieme alle immagini, per ogni elemento passato in input al modello sono state incluse le seguenti tipologie di metadati:

- **Età**: questo attributo indica l'età del paziente. L'età è un fattore significativo nelle diagnosi dermatologiche, poiché alcune patologie cutanee sono più diffuse in gruppi di età specifici.
- **Regione anatomica**: questo metadato indica la localizzazione corporea della lesione cutanea. La prevalenza e le caratteristiche delle lesioni cutanee possono variare in modo significativo a seconda della loro sede anatomica.
- **Diagnosi**: questo attributo fornisce la diagnosi specifica della lesione cutanea, offrendo approfondimenti e categorizzazioni specifiche al di là della classificazione binaria benigna/maligna.
- **Informazioni sulla presenza di melanociti**: indicare se una lesione è melanocitica è fondamentale, soprattutto per riuscire ad identificare al meglio i melanomi.
- **Sesso**: questo attributo è importante perché alcune patologie cutanee possono manifestarsi in modo diverso a seconda del genere del paziente.

3.1.4 Processo di raccolta dei dati

I dati sono stati raccolti attraverso un processo sistematico:

1. **Filtraggio**: il database ISIC è stato consultato per estrarre le immagini che soddisfacevano i requisiti di selezione. Successivamente sono state escluse le immagini con etichette incomplete o ambigue.
2. **Preprocessing dei dati**: i casi indicati come indeterminati/benigni e indeterminati/maligni sono stati standardizzati secondo i criteri definiti per mantenere la coerenza nel set di dati.
3. **Estrazione dei metadati**: i metadati rilevanti per ogni immagine sono stati estratti e utilizzati per la generazione dei prompt associati ad ogni immagine.

3.2 Preprocessing dei dati

Il successo di un modello di apprendimento automatico dipende in larga misura dalla qualità e dal formato dei dati di input. Nel contesto della classificazione delle lesioni cutanee utilizzando i set di dati ISIC, sia le immagini che i metadati associati sono stati sottoposti ad una serie di fasi di pre-elaborazione per garantire una preparazione ottimale per l'addestramento e l'analisi del modello. Questa sezione illustra le metodologie di pre-elaborazione implementate sia per le immagini che per i metadati.

3.2.1 Pre-elaborazione delle immagini

La pre-elaborazione delle immagini è fondamentale per standardizzare i dati di input, migliorare alcune caratteristiche dell'immagine e garantire che il modello riceva i dati in un formato che possa elaborare in modo effieciente. Sono state adottate le seguenti misure:

1. **Etichettatura e pulizia delle immagini**: per garantire la qualità e la pertinenza dei dati, sono state eliminate le immagini prive di etichette chiare che indicassero se la lesione fosse benigna o maligna. Inoltre, le etichette con valori indeterminati, come "indeterminato/benigno" o "indeterminato/maligno", sono state riclassificate assegnandole alla classe benigna e maligna rispettivamente. Questa fase è fondamentale per mantenere la coerenza e l'accuratezza dei dati di addestramento.

2. **Ridimensionamento e normalizzazione delle immagini**: le immagini sono state ridimensionate alla dimensione standard che il modello si aspetta di ricevere in input, in questo caso 224x224 pixel. Inoltre, per ogni immagine è stato eseguito un center crop e la normalizzazione dei valori dei pixel per ogni canale colore in tutto il set di dati. Tutto ciò favorisce la standardizzazione al formato che il modello si aspetta in input e il miglioramento dell'efficienza dell'addestramento.

3.2.2 Pre-elaborazione dei metadati

Oltre alla pre-elaborazione delle immagini, sono stati elaborati anche i metadati associati a ciascuna immagine:

1. **Selezione delle caratteristiche**: in base al contributo che potevano apportare alla classificazione, sono state selezionate feature come l'età, la localizzazione anatomica, lo stato benigno/maligno, la diagnosi, la presenza di melanociti e il sesso. Queste caratteristiche forniscono un contesto prezioso per ogni immagine, aiutando il modello CLIP a comprendere e classificare le lesioni con maggiore precisione.

2. **Gestione dei dati mancanti**: nei casi di metadati incompleti o mancanti, sono state utilizzate strategie come l'imputazione dei dati quando un metadato poteva essere presupposto dalla presenza di un altro metadato, ad esempio nel caso di un melanoma in cui non era indicata la presenza di melanociti, quest'ultima poteva essere facilmente deducibile dalla diagnosi della lesione. In ogni caso, è stata utilizzata una generazione dinamica per la creazione del prompt associato ad ogni immagine in modo da gestire la mancanza di alcune carattestiche in modo automatico.

3.2.3 Generazione dei prompt

Un approccio innovativo nella metodologia è stata la generazione di descrizioni in linguaggio naturale per ogni immagine, integrando i metadati pre-elaborati:

- **Logica di creazione delle didascalie**: è stato implementato un processo di creazione dinamica delle didascalie, in cui la didascalia di ogni immagine è stata generata sulla base dei metadati disponibili. Questo processo ha comportato la creazione di un *template base* per la generazione dei prompt e la successiva personalizzazione con i valori dei metadati associati ad ogni specifica immagine.

- **Influenza sull'allenamento del modello**: queste didascalie, che permettono di integrare informazioni contestuali a quelle visive ottenute dall'immagine, sono state progettate per essere passate in input al modello CLIP, con l'obiettivo di migliorare potenzialmente le prestazioni di classificazione.

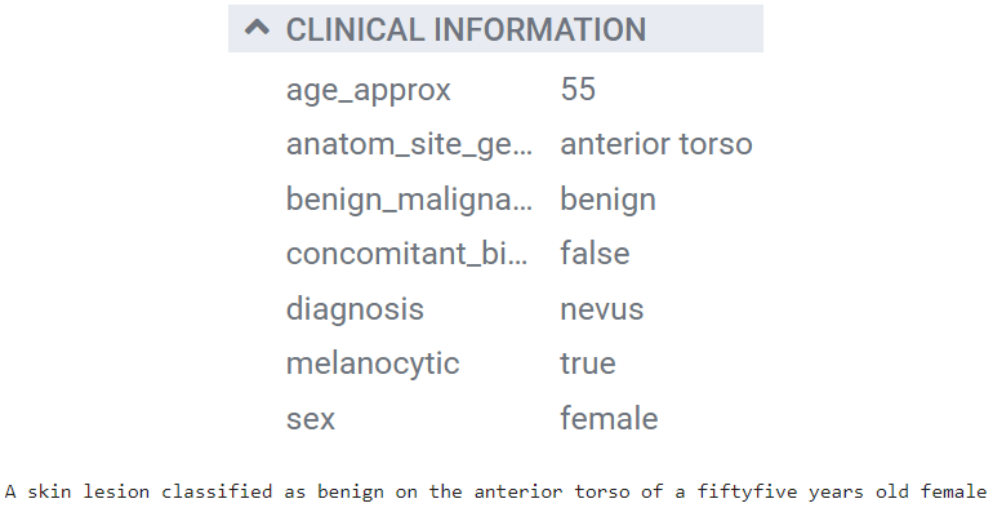

Figura 3.2: Esempio di prompt generato a partire dai metadati disponibili

3.2.4 Integrazione con l'architettura CLIP

La fase finale della pre-elaborazione è stata l'integrazione delle immagini elaborate e delle relative didascalie con il modello CLIP:

- **Input dei dati in CLIP**: è stato utilizzato un approccio sistematico per inserire le immagini pre-elaborate e le relative didascalie generate nel modello CLIP per l'addestramento. Questa fase ha richiesto un'attenta sincronizzazione dei dati dell'immagine e del testo per garantire che ogni immagine fosse abbinata alla sua didascalia corretta.
- **Fine tuning del modello**: nelle prossime sezioni verranno descritti gli adattamenti e le personalizzazioni apportate al modello CLIP standard per il task specifico della classificazione delle lesioni cutanee.

3.3 Architettura del modello

L'obiettivo centrale di questo studio è l'implementazione e il fine tuning dell'architettura CLIP allo scopo di classificare le lesioni cutanee utilizzando immagini e metadati associati. Questa sezione illustra l'architettura di base di CLIP, insieme alle modifiche e agli adattamenti specifici apportati per adattarla agli obiettivi dello studio.

3.3.1 Architettura base di CLIP

CLIP è un modello che unisce il Natural Language Processing (NLP) e la Computer Vision per comprendere immagini e testo in maniera unificata. Descriviamo ora in dettaglio il funzionamento tecnico delle due compone di due blocchi principali:

1. **Encoder di immagini**: questo encoder è responsabile dell'elaborazione delle immagini in ingresso e della loro conversione in image embeddings. Nel nostro caso, l'encoder utilizzato per le immagini è stato un Vision Transformer (ViT). Il Vision Transformer divide un'immagine in patches di dimensioni predefinite, che successivamente vengono linearizzate e convertite in un vettore di embedding. Al fine di conservare le informazioni riguardanti la posizione nello spazio relativa di ciascuna patch, vengono aggiunti dei position embeddings al vettore di embedding, in modo tale da poter identificare in modo univoco la posizione della specifica patch rispetto alla griglia di partenza.
   Ogni patch embedding viene passato attraverso un Multi-Head Self-Attention layer che permette al modello di cogliere il contesto globale e le relazioni tra patches diverse dell'immagine.
   Alla sequenza di patch embeddings viene aggiunto un token chiamato CLS, esso viene utilizzato per aggregare tutte le informazioni provenienti dai patch embeddings.

2. **Encoder di testo**: l'encoder testuale di CLIP è un modello linguistico basato su trasformatori.
   Inizialmente, il testo viene passato in input ad un tokenizzatore che ha il compito di trasformare in token parole o parti di parole. Davanti alla sequnza di token generata, viene aggiunto un token CLS che aggrega le informazioni dell'intera sequenza di token e viene utilizzato per i task di classificazione.
   Successivamente, questi token vengono passati in ingresso al text encoder e transformati in text embeddings. Anche questi embedding vengono fatti passare attraverso un Multi-Head Self-Attention layer, in modo tale che il modello impari a comprendere la semantica del testo in ingresso.

3.3.2 Zero-shot learning

Molti modelli di Computer Vision sono caratterizzati da una conoscenza specifica e mostrano di poter ottenere ottime performance rispetto al task su cui sono stati allenati, ma non riescono bene a generalizzare e a gestire immagini fornite loro in ingresso che non rientrano nel dominio su cui sono stati allenati. Questo significa che per acquisire competenze su un nuovo dominio, diverso da quello su cui è stato allenato il modello, è necessario procedere ad un ulteriore allenamento sul nuovo set di dati. Tutto ciò implica l'utilizzo di tempo per la ricerca di nuove immagini appartenenti al nuovo dominio di interesse e il successivo allenamento sul nuovo set di dati.

Modelli che necessitano di diverse immagini per specializzarsi in un task specifico sono chiamati N-shot Learners, in quanto hanno bisogno che vengano forniti loro in ingresso N esempi di immagini per poter procedere alla classificazione. Idealmente, sarebbe preferibile che un modello non necessitasse di alcun esempio in fase di training e che comunque fosse in grado di ottenere delle performance elevate.

CLIP è stato pre-allenato da OpenAI usando un dataset di 400 milioni di coppie immagine-testo disponibili su internet. Questo ha permesso a CLIP di ottenere una conoscenza solida del contesto generico di una frase associata ad un'immagine e non solo dell'etichetta specifica relativa all'immagine. Grazie alla grande quantità di dati su cui è stato allenato, CLIP mostra ottime performance nella classificazione di immagini in zero-shot, ovvero non dovendo riallenare il modello usando immagini specifiche per il task di classificazione di interesse.

In seguito alla codifica dell'immagine e del testo ad opera rispettivamente dell'image encoder e del text encoder, gli embedding generati avranno la dimensione di un vettore di 512 elementi. Essi saranno poi proiettati in uno spazio vettoriale ad alta dimensionalità e, per valutare la similarità tra l'embedding dell'immagine e quello testuale verrà calcolato un coefficiente di allineamento sfruttando la cosine similarity o il dot product. La coppia immagine-testo che restituirà il coefficiente di similarità più alto sarà quella che il modello pensa possa essere più altamente correlata.

Nell'esempio presente di seguito, passando in ingresso a CLIP l'immagine di un gatto che non è stata fornita per l'allenamento del modello, e come descrizione testuale "A photo of a car/bird/cat", il calcolo della cosine similarity restituirà un valore più alto associato alla frase "A photo of a cat".

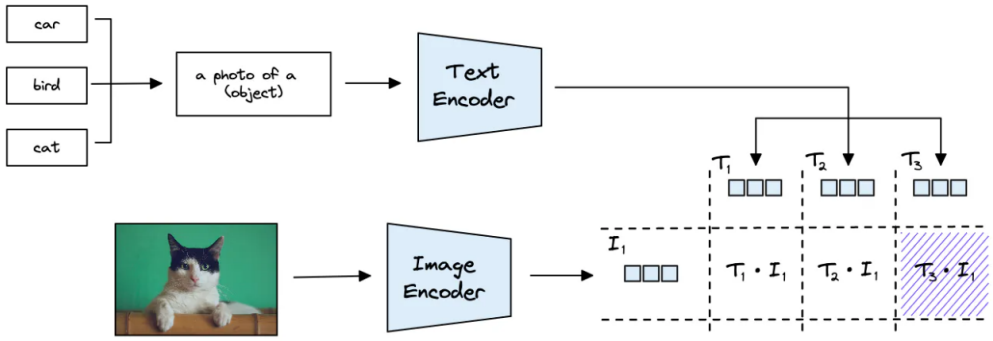

Figura 3.3: Esempio di calcolo del coefficiente di similarità [22]

3.3.3 Contrastive Learning

Durante il training, sia gli embeddings delle immagini che quelli dei testi vengono proiettati in uno stesso spazio vettoriale multidimensionale. I due encoder sono simultaneamente addestrati con l'obiettivo di allineare gli embeddings corrispondenti e di separare quelli non corrispondenti.

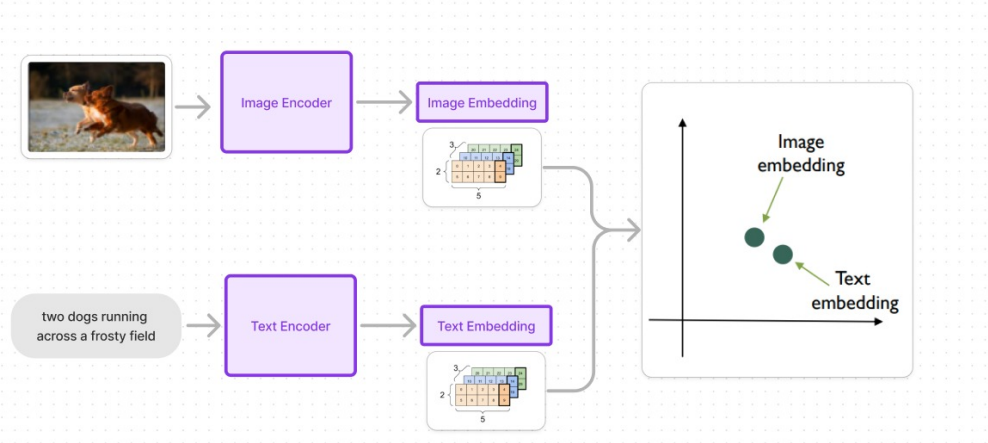

Figura 3.4: Il testo e l'immagine vengono proiettati in un nuovo iperspazio in seguito all'encoding

Più specificamente, l'obiettivo dell'addestramento è massimizzare la cosine similarity tra gli embeddings dell'immagine $\mathbf{v}_i$ e le corrispondenti descrizioni testuali $\mathbf{u}_i$:

$$\max_{\theta} \mathbb{E}_{(i,j)\sim\mathcal{P}} \left[\frac{\mathbf{v}_i^\top \mathbf{u}_i}{\|\mathbf{v}_i\| \|\mathbf{u}_i\|} \right]$$

dove θ rappresenta i parametri degli encoder, e $\mathcal{P}$ è la distribuzione di coppie corrispondenti immagine-testo.

Contemporaneamente, si minimizza la cosine similarity tra un'immagine $\mathbf{v}_i$ e le descrizioni testuali non corrispondenti $\mathbf{u}_j$:

$$\min_{\theta} \mathbb{E}_{(i,j)\sim\mathcal{Q}} \left[\frac{\mathbf{v}_i^\top \mathbf{u}_j}{\|\mathbf{v}_i\| \|\mathbf{u}_j\|} \right]$$

dove $\mathcal{Q}$ è la distribuzione di coppie non corrispondenti.

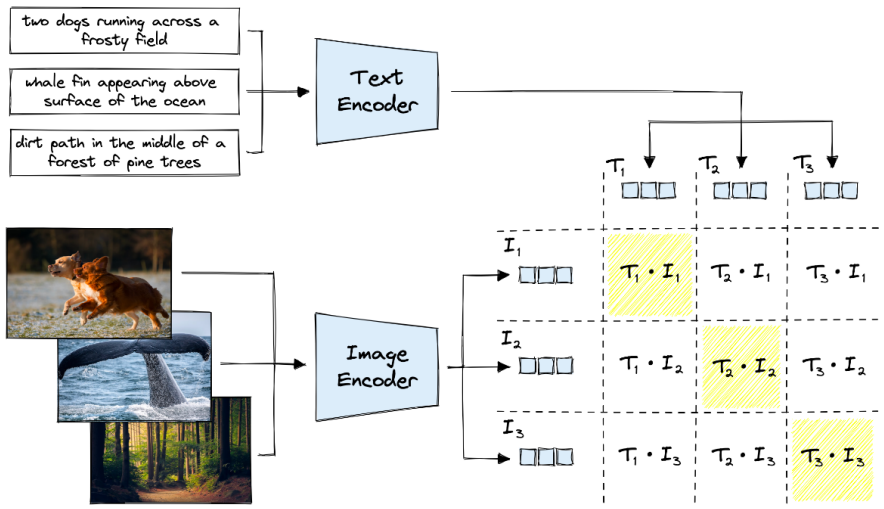

Figura 3.5: Meccanismo di Contrastive Learning

In sintesi, il Contrastive Learning mira ad ottimizzare gli encoder in modo che gli embeddings corrispondenti siano vicini (massimizzando la cosine similarity) e gli embeddings non corrispondenti siano distanti (minimizzando la cosine similarity) nello spazio vettoriale condiviso. Questo approccio favorisce la creazione di rappresentazioni semanticamente ricche e utili per applicazioni di matching immagine-testo e altri task di elaborazione del linguaggio e visione artificiale.

3.3.4 Scelta del modello

Ad oggi, OpenAI mette a disposizione diversi modelli CLIP. Tra questi ci sono:

- **ViT-B/32**: il Vision Transformer utilizzato si aspetta in ingresso un'immagine di dimensioni 224x224 pixel, suddivisa in 32x32 patches, per un totale di 1024 patches ricavate dalla singola immagine di partenza. Ogni patch ha dimensione di 7x7 pixel.

- **ViT-B/16**: il Vision Transformer prende come input immagini di 224x224 pixel, suddivise in 16x16 patches, per un totale di 256 patches. Ogni patch ha dimensione 14x14 pixel.
- **ViT-L/14**: questo Vision Transformer si aspetta in ingresso un'immagine di 224x224 pixel, suddivisa in 14x14 patches, per un totale di 196 patches. Ogni patch ha dimensione di 16x16 pixel.
- **ViT-L/14@336px**: in questo caso il modello prende in ingresso immagini con una risoluzione più alta pari a 336x336 pixel. Ogni immagine viene suddivisa in 14x14 patches e ogni patch risulta avere dimensione di 24x24 pixel.

Il modello con il quale abbiamo deciso di procedere per effettuare il fine tuning è il ViT-B/32, in quanto risulta essere quello che necessita di meno risorse computazionali. Gli attention layer presenti all'interno del Vision Transformer richiedono che venga effettuato il confronto tra ogni input e tutti gli altri possibili input, quindi nel caso del ViT-B/32 ogni attention layer deve gestire una quantità di circa 2.5 milioni di confronti. Per gli altri modelli, la quantità di confronti da effettuare aumenta passando dai circa 9.8 milioni per il modello ViT-B/16, ai circa 65 milioni per il ViT-L/14@336px.

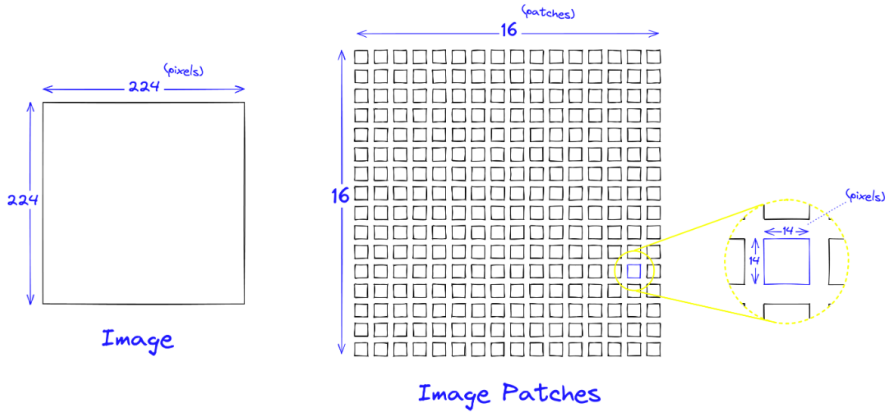

Figura 3.6: Divisione in patches per il modello ViT-B/16

3.4 Tuning Pipeline

Per il task specifico della classificazione delle lesioni cutanee, sono state apportate le seguenti modifiche all'architettura di CLIP:

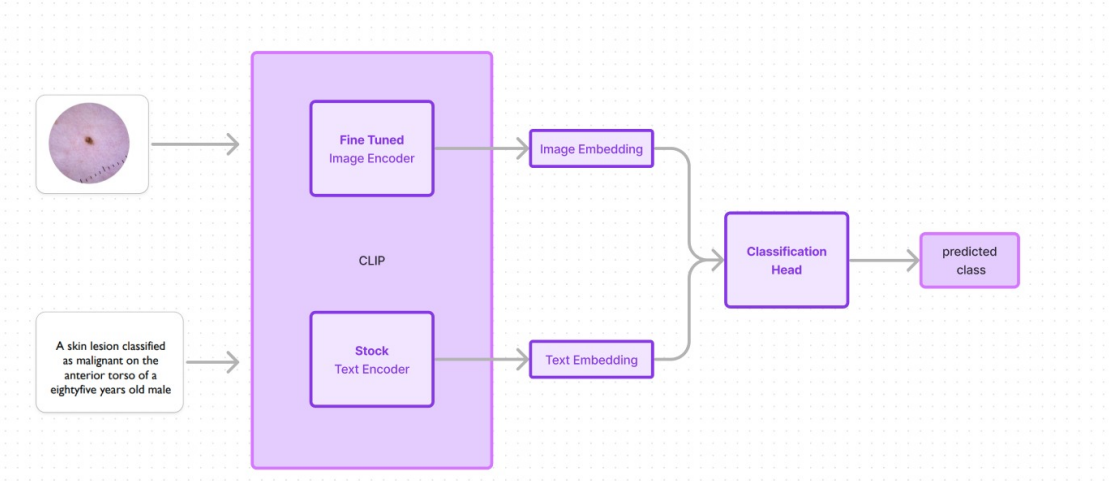

Figura 3.7: Fine tuning del modello

3.4.1 Fine tuning dell'image encoder

Dato il task altamente specifico del riconoscimento delle lesioni cutanee, ciò che è stato fatto è stato riallenare il Vision Transformer passando in ingresso le immagini delle lesioni cutanee.

Le immagini sono state raggruppate in batch, in modo tale da passare in input al modello diverse coppie immagine-testo. Un punto critico dello studio è stato trovare il numero più adatto di immagini da inserire in un batch, in quanto l'obiettivo di CLIP è utilizzare il Contrastive Learning per far avvicinare gli embedding relativi a immagine-testo associati e far allontanare le coppie immagine-testo non corrispondenti. Ciò presupponeva il fatto che nello stesso batch non dovessero esserci immagini con una descrizione testuale simile, in quanto questo poteva comportare confusione da parte del modello.

Il batch contenente le immagini e i relativi prompt vengono passati in input al modello, che ne calcola in modo separato gli image embedding e i text embedding. Il modello a questo punto calcola i coefficienti di similarità per ogni immagine del batch rispetto ad ogni descrizione testuale presente all'interno del batch e la stessa cosa viene fatta al contrario, ovvero viene calcolata la similarità per ogni

testo del batch rispetto ad ogni immagine. Ciò che viene fatto è quindi andare a valutare quanto una certa immagine sia più facilmente associabile ad una serie di descrizioni testuali e viceversa.

La funzione di loss viene calcolata come la media delle loss derivanti dal calcolo della similarità di ogni immagine per tutti i testi del batch e dalla similarità di ogni descrizione testuale a tutte le immagini del batch, usando la funzione CrossEntropy di Pytorch. L'errore calcolato è quello che viene retropropagato per il successivo aggiornamento dei pesi del modello, con l'obiettivo di avvicinare gli accoppiamenti corretti tra immagine e testo nello spazio vettoriale multidimensionale e allontanare coppie immagine-testo non associate.

Le performance ottenute con la variazione dei pesi sono state valutate sul development set, in modo tale da verificare come il cambiamento dei pesi della rete influisse su un set di dati su cui il modello non era stato allenato. Al fine di osservare un miglioramento delle performance del modello sul development set si è scelto di salvare i pesi del modello associati al valore di loss più basso ottenuto sul development set e di arrestare l'allenamento del modello in caso di aumento della loss per più di tre epoche consecutive.

3.4.2 Ottimizzazione degli iperparametri

Al fine di ottimizzare le performance del modello per lo specifico task di classificazione, durante il fine tuning sono stati regolati diversi iperparametri:

- **Learning rate**: è il parametro tramite il quale si seleziona il passo che l'algoritmo di ottimizzazione deve compiere al termine di ogni iterazione. Questo parametro contribuisce alla modifica dei pesi del modello e la sua ottimizzazione risulta di fondamentale importanza, in quanto valori di learning rate troppo alti potrebbero causare una mancata convergenza mentre valori troppo bassi rallenterebbero il processo di allenamento della rete e potrebbero portare ad una convergenza in un minimo locale. In questo studio sono state analizzate le performance ottenute andando a variare il valore del learning rate ed è stato selezionato 1e−5.

- **Dimensione del batch**: questo parametro indica il numero di coppie immagine-didascalia che devono essere passate in input al modello durante un'iterazione. Batch di dimensione più grande possono portare a raggiungere una convergenza più stabile ma richiedono di avere a disposizione risorse computazionali notevoli, mentre un batch size piccolo porta ad un aggiornamento più frequente dei pesi, ma allo stesso tempo è più soggetto al rumore. In questo caso si è scelto di procedere utilizzando un batch size pari a 64.

Usando CLIP, la dimensione del batch risulta essere di fondamentale importanza, in quanto il meccanismo di Contrastive Learning si basa sulla necessità di avere un unico accoppiamento corretto tra immagine e testo. Quindi, selezionando un batch size troppo grande, il rischio che nello stesso batch siano presenti immagini con la stessa descrizione testuale aumenta.

- **Numero di epoche**: indica il numero di passaggi completi attraverso tutti gli elementi del training set. Un valore più elevato di epoche potrebbe permettere al modello di convergere in un minimo globale ma aumenta anche il rischio di overfitting.
- **Ottimizzatore**: serve per aggiornare i pesi del modello durante l'allenamento. Nel nostro caso, si è deciso di utilizzare Adam come algoritmo di ottimizzazione.
- **Learning rate scheduler**: questo parametro permette di regolare il valore del learning rate durante le epoche di allenamento. Nel progetto è stato scelto che ogni tre epoche il learning rate venisse ridotto di un decimo.
- **Early stopping**: questo parametro serve per prevenire l'overfitting sul training set quando le performance sul development set smettono di migliorare. In particolare, è possibile impostare il numero di epoche che si possono aspettare in attesa che si osservi un miglioramento nella loss del development set prima di arrestare il training. In questo specifico caso, il numero di epoche di attesa è stato settato a tre.

3.4.3 Aggiunta di un classification head

In successione al modello fine tuned è stato aggiunto un ulteriore layer di classificazione. Il classification head è composto da un layer fully connected che applica una trasformazione lineare agli embedding calcolati per l'immagine e la rispettiva descrizione testuale. In seguito alla trasformazione lineare vengono poi determinate le predizioni associate ad ogni input e viene calcolato un errore utilizzando la binary crossentropy sulla base della predizione e della label riferita a quell'input. Questo errore viene retropropagato e i pesi interni del classification head vengono aggiornati. Inoltre, alla funzione di binary crossentropy è stato aggiunto un ulteriore parametro per tenere conto della class imbalance tra lesioni maligne e lesioni benigne. In particolare, un errore fatto nella classificazione di una lesione maligna risulta pesare di più rispetto ad un errore sulla classe benigna.

Queste modifiche all'architettura di CLIP sono state apportate per ottimizzare le prestazioni del modello per il task specifico della classificazione delle lesioni cutanee, assicurandosi che potesse sfruttare al meglio le informazioni ricavabili dalla combinazione di contesti visivi e testuali.

3.5 Metriche di valutazione

Per l'analisi del sistema di classificazione delle lesioni cutanee basato su CLIP, sono state utilizzate diverse metriche chiave per valutare le prestazioni del modello in modo completo. Queste metriche sono state scelte per la loro rilevanza nel contesto della classificazione delle immagini mediche e per la loro capacità di fornire una visione più chiara delle capacità del modello:

- **Matrice di confusione**: si tratta di uno strumento fondamentale nelle attività di classificazione, in quanto fornisce una ripartizione dettagliata delle prestazioni del modello tra le diverse classi. Mostra il numero di veri positivi, veri negativi, falsi positivi e falsi negativi, offrendo una visuale completa sul tipo di errori che il modello sta commettendo.

- **Curva ROC**: questa curva è un indicatore delle performance di classificazione a diversi livelli di soglia. Ciò che serve calcolare è il True Positive Rate (TPR), chiamato anche sensibilità, e il False Positive Rate (FPR), ovvero il rapporto tra gli elementi classificati incorrettamente come positivi e la somma di tutti i campioni negativi. Questa curva permette di osservare il trade-off tra sensibilità e specificità, inoltre l'Area Under the Curve (AUC) è una misura indicativa delle performance del modello per tutti i possibili valori di soglia.

- **Accuratezza**: questa metrica misura la percentuale di esempi classificati correttamente rispetto al numero totale di istanze. Fornisce un'idea generale delle prestazioni del modello, ma non è la scelta ideale per valutare le performance per set di dati sbilanciati.

- **Sensibilità (recall)**: la sensibilità misura la percentuale di casi effettivamente positivi identificati correttamente dal modello. Nell'imaging medico, questo è fondamentale per identificare tutti i casi potenziali di una patologia, come nel caso delle lesioni maligne.

- **Specificità**: misura la percentuale di casi effettivamente negativi identificati correttamente. Nel contesto della classificazione delle lesioni cutanee, riflette la capacità del modello di identificare correttamente le lesioni benigne.

- **Precisione**: la precisione è il rapporto tra i veri positivi e la somma dei veri e dei falsi positivi. Indica l'accuratezza delle previsioni positive, un fattore essenziale nelle diagnosi mediche, dove i falsi allarmi possono portare a trattamenti non necessari.

- **F1 score**: il punteggio F1 è la media armonica di precisione e recall. Si tratta di un'unica metrica che bilancia sia la precisione che la recall, fornendo una

visione più completa delle prestazioni del modello, soprattutto in set di dati in cui lo sbilanciamento delle classi è importante.

Ciascuna di queste metriche è stata scelta per garantire una valutazione approfondita del modello. Nel contesto della tesi, queste metriche non devono essere solo calcolate ma anche interpretate alla luce del loro significato nella classificazione delle immagini mediche. Ciò consente di comprendere chiaramente le prestazioni del modello e gli eventuali miglioramenti da apportare, soprattutto in termini di applicabilità clinica.

Capitolo 4

Risultati

4.1 Meccanismo di testing

4.1.1 Zero-shot CLIP

Il meccanismo di testing in zero-shot è stato organizzato nel modo seguente:

- **Generazione di due prompt**: sono state create due strutture di descrizioni testuali "A benign skin lesion" e "A malignant skin lesion".

- **Tokenizzazione e creazione dei text embeddings**: i due prompt sono stati successivamente transformati in token tramite il tokenizzatore e trasformati in text embeddings usando il text encoder.

- **Pre-elaborazione delle immagini e creazione degli image embeddings**: le immagini sono state preprocessate e passate in input al modello, a seguire sono state trasformate in image embeddings tramite l'azione dell'image encoder.

- **Calcolo della similarità**: è stata calcolata la cosine similarity tra gli image embeddings e i due text embeddings riferiti ad ognuna delle due classi. La classe predetta associata ad ogni immagine è stata stabilita usando la funzione softmax, che permette di calcolare la probabilità che a quell'immagine sia associata ognuna delle due descrizioni testuali.

- **Confronto con la classe reale**: infine le predizioni sono state confrontate con la classe reale associata ad ogni immagine e sono state calcolate le metriche di valutazione.

4.1.2 Modello fine tuned con classification head

La fase di testing per il modello fine tuned con classification head è stata implementata nelle seguenti fasi:

- **Generazione del prompt**: partendo dai metadati disponibili per la specifica immagine, è stata generata una descrizione testuale basandosi sul template "A skin lesion on the [anatomy] of a [age] years old [sex]", dove i termini tra parantesi quadre sono stati sostituiti con i metadati associati all'immagine. Non sono stati considerati in fase di testing i metadati che indicavano la diagnosi della lesione e la presenza o l'assenza di melanociti, in quanto risultano essere dei metadati che hanno un ruolo significativo nella determinazione della benignità o malignità di una lesione cutanea.
- **Tokenizzazione e creazione dei text embeddings**: il prompt è stato transformato in token tramite il tokenizzatore e successivamente convertito in text embeddings usando il text encoder.
- **Pre-elaborazione delle immagini e creazione degli image embeddings**: le immagini sono state preprocessate e passate in input al modello, a seguire sono state trasformate in image embeddings tramite l'azione dell'image encoder.
- **Calcolo delle predizioni**: sono state calcolate le probabilità utilizzando una funzione sigmoide e fissando una soglia pari a 0.5, per probabilità al di sotto di questa soglia la classe predetta risulta essere benigna, mentre al di sopra della soglia la classe predetta è maligna.
- **Confronto con la classe reale**: infine le predizioni sono state confrontate con la classe reale associata ad ogni immagine e sono state calcolate le metriche di valutazione.

4.2 Performance del modello

4.2.1 Zero-shot classification

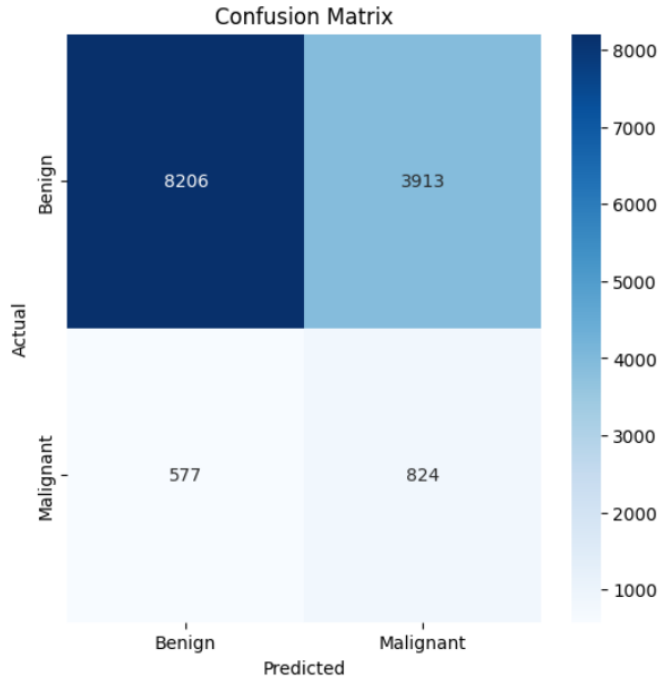

Figura 4.1: Confusion Matrix di tutto il set di dati in zero-shot

In questo primo studio, è stata investigata l'efficacia del modello CLIP nella classificazione zero-shot di immagini di lesioni cutanee, distinguendo tra lesioni benigne e maligne. I risultati ottenuti sul set di dati mostrano performance differenti nelle metriche di valutazione:

- L'**accuratezza** generale del modello è risultata del **67%**, suggerendo una capacità moderata di classificazione complessiva.
- La **sensibilità** del modello è pari al **59%**, il che indica che il modello classifica correttamente come maligni il 59% dei casi totali di lesioni maligne.
- La **specificità** del modello è risultata del **68%**, il che indica che il modello classifica correttamente come benigni il 68% dei casi totali di lesioni benigne.
- La **precisione** per la **classe benigna** è risultata elevata, con un valore del **93%**, indicando che quando il modello predice un'immagine come benigna, essa è realmente benigna il 93% delle volte.

- Tuttavia, la **precisione** per la **classe maligna** è stata notevolmente inferiore, pari al **17%**, il che indica una tendenza del modello a identificare erroneamente molte lesioni cutanee benigne come maligne.
- L'**F1-score**, metrica che bilancia precisione e recall, ha mostrato un valore del **79%** per la **classe benigna** e del **27%** per la **classe maligna**, sottolineando una maggiore difficoltà del modello nel rilevare correttamente le lesioni cutanee maligne.

I fattori che potrebbero aver influenzato questi risultati comprendono la complessità e la varietà delle lesioni cutanee presenti nel dataset, nonché la capacità del modello CLIP di generalizzare efficacemente su immagini su cui non è stato esplicitamente addestrato durante la fase di training.

In conclusione, sebbene il modello CLIP abbia dimostrato capacità promettenti nella classificazione zero-shot di immagini generiche, la sua applicazione specifica alle immagini di lesioni cutanee ha evidenziato limitazioni significative, soprattutto nella discriminazione accurata delle lesioni maligne.

4.2.2 CLIP fine tuned

L'applicazione del modello CLIP fine tuned al dataset ISIC per la classificazione di immagini dermatologiche in benigne e maligne ha prodotto risultati promettenti, pur evidenziando alcune aree di miglioramento.

Sul training set, il modello ha dimostrato di raggiungere un'accuratezza complessiva del 91.8%, indicando una buona capacità di apprendimento dai dati forniti. Particolarmente notevole è stata la performance nel rilevamento dei casi maligni, con una recall del 99.4%, suggerendo un'elevata sensibilità del modello nell'identificare potenziali melanomi.

Figura 4.2: Confusion Matrix sul training set usando il modello CLIP fine tuned

Tuttavia, l'analisi del development set ha rivelato alcune criticità. L'accuratezza è scesa all'81.5%, segnalando un certo grado di overfitting. La precisione nella classificazione delle immagini maligne si è attestata al 39.1%, indicando una tendenza del modello a sovrastimare i casi positivi. Nonostante ciò, il modello ha mantenuto una buona recall (84.6%) anche sul development set, confermando la sua capacità di identificare la maggior parte dei casi maligni correttamente.

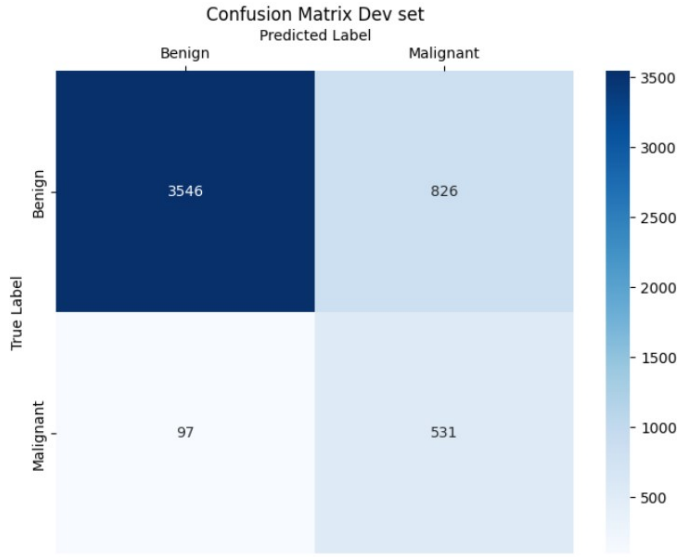

Figura 4.3: Confusion Matrix sul development set usando il modello CLIP fine tuned

La specificità per le immagini benigne si è mantenuta relativamente alta sia nel training (90.7%) che nel development set (81.1%), dimostrando una buona abilità nel riconoscere correttamente i casi non patologici.

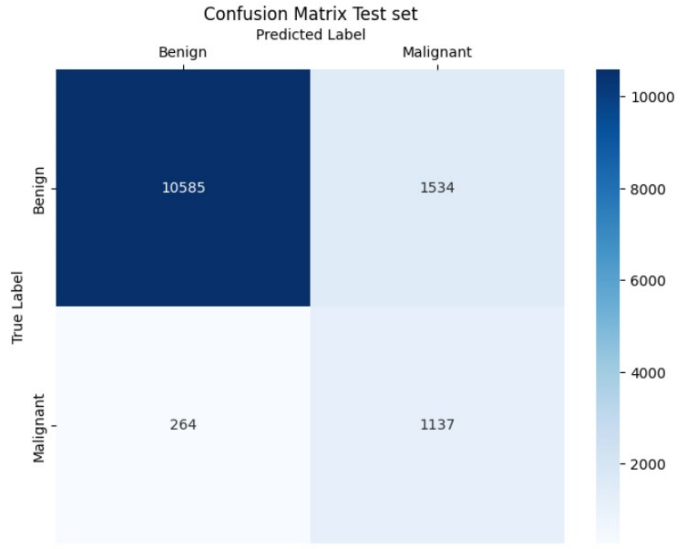

Figura 4.4: Confusion Matrix sul test set usando il modello CLIP fine tuned

Le metriche di valutazione sul test set con il modello CLIP fine tuned mostrano i seguenti risultati:

- L'**accuratezza** generale del modello è risultata dell'**87%**, mostrando una capacità elevata di classificazione complessiva.
- La **sensibilità** del modello è pari all'**81%**, il che indica che il modello classifica correttamente come maligni l'81% dei casi totali di lesioni maligne.
- La **specificità** del modello è risultata dell'**87%**, il che indica che il modello classifica correttamente come benigni l'87% dei casi totali di lesioni benigne.
- La **precisione** per la **classe benigna** è risultata elevata, con un valore del **98%**, indicando che quando il modello predice un'immagine come benigna, essa è realmente benigna il 98% delle volte.
- Tuttavia, la **precisione** per la **classe maligna** è stata moderatamente inferiore, pari al **43%**, il che indica una tendenza del modello a identificare erroneamente molte lesioni cutanee benigne come maligne.
- L'**F1-score**, metrica che bilancia precisione e recall, ha mostrato un valore del **92%** per la **classe benigna** e del **56%** per la **classe maligna**, sottolineando una maggiore difficoltà del modello nel rilevare correttamente le lesioni cutanee maligne.

Questi risultati suggeriscono che il modello CLIP fine tuned ha potenziale nell'assistere la diagnosi dermatologica, soprattutto come strumento di screening iniziale per identificare casi sospetti che richiedono ulteriori indagini. La sua alta sensibilità lo rende particolarmente utile in contesti dove è cruciale minimizzare i falsi negativi.

Tuttavia, l'elevato numero di falsi positivi, specialmente nel development set e nel test set, indica la necessità di ulteriori modifiche. Nella prossima sezione andremo ad analizzare le performance ottenute del modello in seguito all'aggiunta del classification head.

4.2.3 Risultati con classification head

Utilizzando il modello CLIP fine tuned e aggiungendo in coda il classification head addestrato sul set di dati di training e validato sul development set, i risultati ottenuti sono quelli mostrati in seguito.

Sul training set, il modello complessivo ha dimostrato di ottenere un valore di accuratezza pari al 98%, mostrando un considerevole aumento delle prestazioni rispetto al modello con solo CLIP fine tuned. La recall per la classe maligna invece risulta essere peggiorata lievemente, arrivando al valore di 96%, indicando che il modello tende a classificare più immagini maligne come benigne. Allo stesso tempo però, la precisione sulla classe maligna è passata dal 61% del modello con CLIP fine tuned al 91% del modello con l'aggiunta del classification head, il che significa che il numero di falsi positivi si è considerevolmente ridotto.

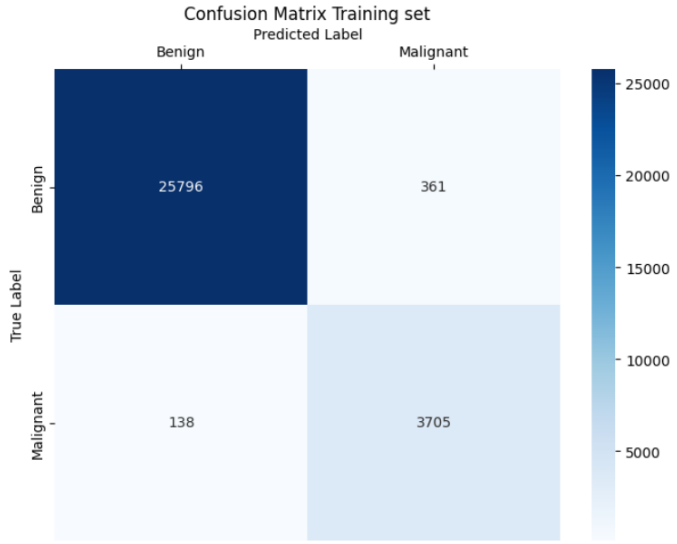

Figura 4.5: Confusion Matrix sul training set con il classification head

I risultati sul development set mostrano come l'accuratezza sia rimasta comunque elevata, pari al 93%, mentre la precisione nella classificazione delle lesioni maligne è salita al 69% rispetto al 39% del modello CLIP fine tuned. La recall per la classe maligna invece risulta essere leggermente peggiorata rispetto al modello CLIP fine tuned, arrivando ad un valore del 74%, indicando che il modello tende a classificare come benigne più immagini che in realtà sono maligne.

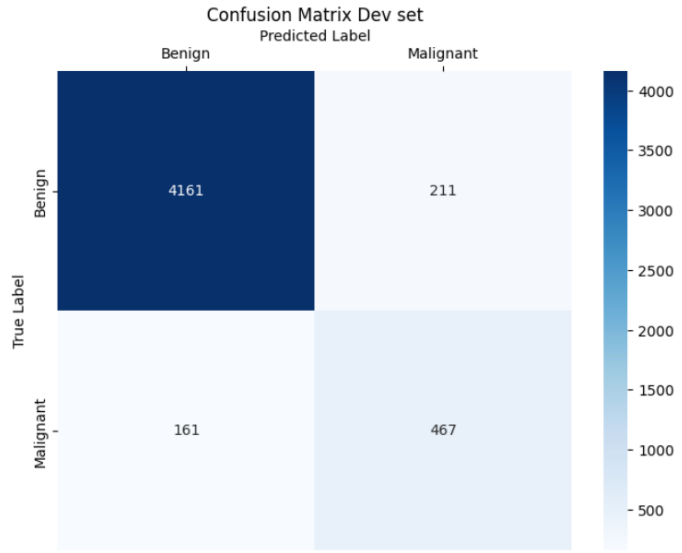

Figura 4.6: Confusion Matrix sul development set con il classification head

La specificità si è mantenuta alta sia nel training set, con un valore pari al 99%, che nel development set, in cui risulta pari al 95%, mostrando un'ottima capacità nella corretta classificazione delle lesioni benigne.

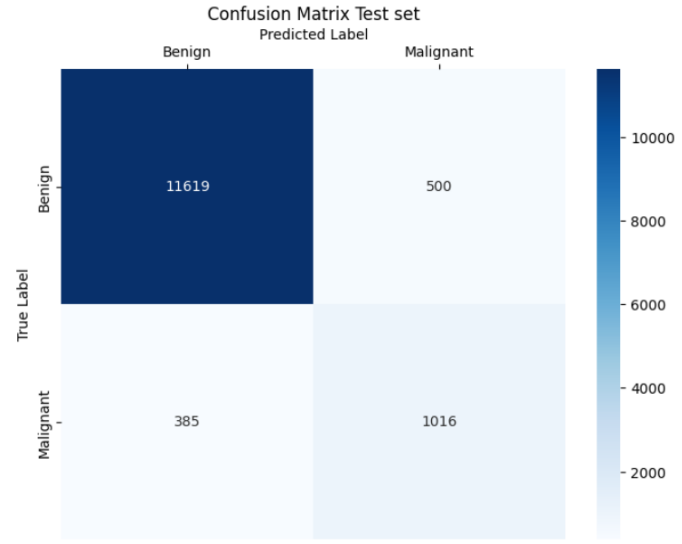

Figura 4.7: Confusion Matrix sul test set con il classification head

Le metriche di valutazione sul test set con il modello CLIP fine tuned e il classification head mostrano i seguenti risultati:

- L'**accuratezza** generale del modello è risultata dell'**93%**, mostrando una capacità elevata di classificazione complessiva.
- La **sensibilità** del modello è pari al **73%**, il che indica che il modello classifica correttamente come maligni il 73% dei casi totali di lesioni maligne.
- La **specificità** del modello è risultata del **96%**, il che indica che il modello classifica correttamente come benigni il 96% dei casi totali di lesioni benigne.
- La **precisione** per la **classe benigna** è risultata elevata, con un valore del **97%**, indicando che quando il modello predice un'immagine come benigna, essa è realmente benigna il 97% delle volte.
- La **precisione** per la **classe maligna** è stata discretamente inferiore, pari al **67%**, il che indica una tendenza del modello a identificare erroneamente alcune lesioni cutanee benigne come maligne.
- L'**F1-score**, metrica che bilancia precisione e recall, ha mostrato un valore del **96%** per la **classe benigna** e del **70%** per la **classe maligna**, sottolineando una maggiore difficoltà del modello nel rilevare correttamente le lesioni cutanee maligne.

Il modello con l'aggiunta del classification head ha ottenuto delle buone performance in termini di accuratezza e specificità, la precisione per la classe maligna e la recall invece hanno mostrato delle prestazioni moderate. Inoltre, l'F1-score indica che si è raggiunto un buon compromesso tra la precisione e la recall per la classe maligna.

4.3 Confronto con altri modelli

	Gessert	Ningrum	Jasil	FT CLIP	DeepMetaForge
Accuracy	92%	98.5%	98.2%	93%	99.5%
Precision (Benign)	-	-	-	97%	99.7%
Precision (Malignant)	-	-	-	67%	-
Recall (Benign)	92%	100%	100%	96%	85%
Recall (Malignant)	93%	-	-	73%	-
F1-score (Benign)	96%	99.2%	99.1%	96%	99.7%
F1-score (Malignant)	33%	-	-	70%	-
AUC	-	-	-	95%	84%

Tabella 4.1: Performance ottenute con diversi modelli

Il modello CLIP fine tuned è stato confrontato con altri modelli presenti nel paper DeepMetaForge [23]. In particolare, Gessert [24] analizza le performance ottenute da una rete di tipo EfficientNets, Ningrum [25] mostra le performance relative ad una rete CNN affiancata da un'ANN per la gestione dei metadati, Jasil [26] utilizza un approccio con una CNN ibrida.
Dall'osservazione della tabella si possono fare le seguenti considerazioni:

- L'accuratezza del modello CLIP fine tuned risulta essere più elevata del modello di Gessert, ma più bassa degli altri modelli. In generale però, tutte le accuratezze riportate mostrano dei valori molto elevati e paragonabili.
- La precisione per la classe benigna è indicata solo per il modello DeepMetaForge e risulta essere paragonabile a quella raggiunta dal modello CLIP fine tuned.
- La precisione per la classe maligna non è indicata per gli altri modelli, nel caso di CLIP fine tuned questo risultato ci dice che il numero di veri positivi rispetto alla somma di tutte le lesioni classificate dal modello come maligne è pari al 67%. Questo valore sembrerebbe essere dovuto al grande sbilanciamento delle classi presenti nel dataset, in sviluppi futuri si potrebbe provare ad allenare un modello con un dataset in cui le lesioni della classe benigna e della classe maligna sono perfettamente bilanciate.
- La recall per la classe benigna mostra risultati migliori rispetto a due dei modelli presenti in tabella e paragonabili agli alti risultati raggiunti dagli altri modelli.
- La recall per la classe maligna risulta essere moderatamente inferiore rispetto a quella presente nel modello di Gessert, indicando quindi la presenza di una percentuale più alta di falsi negativi, e quindi di immagini di lesioni maligne classificate erroneamente come benigne, nel modello CLIP fine tuned.

- L'F1-score per la classe benigna mostra prestazioni paragonabili a quelle degli altri modelli.
- L'F1-score per la classe maligna risulta essere molto più elevata rispetto al modello di Gessert, andando ad indicare che il modello CLIP fine tuned mostra un buon bilanciamento tra precisione e recall per la classe maligna.
- L'AUC della curva ROC del modello CLIP fine tuned può essere confrontata solo al modello DeepMetaForge e mostra dei valori significativamente migliori.

Capitolo 5

Conclusioni

I risultati sul test set indicano che il modello CLIP fine-tuned ha dimostrato una buona performance generale nella classificazione delle lesioni cutanee in benigne e maligne. Tuttavia, la precisione e il recall per la classe maligna rimangono moderati, suggerendo la necessità di ulteriori miglioramenti per ridurre il numero di falsi positivi e falsi negativi.

Durante lo sviluppo del modello, è emersa una complicazione significativa dovuta all'assenza di molti metadati. Molte informazioni erano mancanti o contrassegnate come "NA", limitando così la capacità del modello di apprendere completamente dalle correlazioni tra immagini e dati testuali.

Per affrontare queste sfide, sebbene siano già state adottate tecniche come l'augmentation delle immagini, il dropout sul classification head e la gestione del class imbalance con augmentation e weighted class approach, è possibile migliorare ulteriormente l'applicazione di queste tecniche o esplorare modalità alternative:

- **Ottimizzazione dell'Augmentation delle Immagini**: L'augmentation delle immagini può essere ulteriormente raffinata, sperimentando con diverse tecniche di trasformazione, come rotazioni, traslazioni, variazioni di luminosità e contrasto, per migliorare la robustezza del modello.
- **Dropout**: Sebbene sia stato applicato dropout sul classification head, potrebbe essere utile esplorare l'implementazione di dropout a livelli diversi del modello o variare la probabilità di dropout per trovare un equilibrio ottimale che riduca l'overfitting.
- **Prompt Engineering**: L'importanza del dataset dei prompt è cruciale. Migliorare la qualità e la diversità dei prompt testuali utilizzati potrebbe ottimizzare l'apprendimento delle correlazioni tra immagini e descrizioni testuali, potenziando la capacità del modello di distinguere tra lesioni benigne e maligne.

- **Gestione del Class Imbalance**: Nonostante l'uso di tecniche di augmentation e approcci weighted class, si potrebbero sperimentare ulteriori strategie, come il focal loss o la sintesi di nuovi dati per le classi minoritarie, al fine di migliorare l'equilibrio tra le classi.

Implicazioni Cliniche e Futuri Sviluppi

L'utilizzo del modello CLIP per la classificazione automatica delle lesioni cutanee rappresenta un significativo passo avanti nella diagnosi precoce del cancro alla pelle. Questo approccio multimodale supporta i dermatologi nella valutazione delle lesioni cutanee, riducendo il carico di lavoro e permettendo diagnosi più tempestive e accurate.

Le potenzialità di CLIP aprono nuove prospettive di ricerca nel campo dell'intelligenza artificiale applicata alla dermatologia. Future ricerche potrebbero esplorare:

- **Integrazione con Sistemi di Supporto Decisionale Clinico (CDSS)**: Lo sviluppo di sistemi che integrano CLIP con altre fonti di dati clinici potrebbe fornire diagnosi ancora più accurate e personalizzate.
- **Valutazione Longitudinale**: Studi che valutano le performance del modello nel tempo, considerando l'evoluzione delle lesioni cutanee e l'impatto delle diverse terapie, potrebbero offrire ulteriori insights sull'efficacia clinica di questo approccio.
- **Collaborazioni Interdisciplinari**: Collaborazioni tra esperti di intelligenza artificiale, dermatologi e altri specialisti medici potrebbero portare a innovazioni significative, migliorando continuamente l'accuratezza e l'affidabilità dei modelli diagnostici.

In conclusione, questo studio dimostra che l'impiego del modello CLIP nella classificazione delle lesioni cutanee può rivoluzionare la metodologia diagnostica, con notevoli benefici per la salute pubblica. Il continuo sviluppo e miglioramento di tali modelli rappresenta una promettente frontiera nella lotta contro il cancro alla pelle e altre patologie dermatologiche.

Bibliografia

[1] Felix Krones, Umar Marikkar, Guy Parsons, Adam Szmul e Adam Mahdi. «Review of multimodal machine learning approaches in healthcare». In: *ArXiv* abs/2402.02460 (2024). URL: https://api.semanticscholar.org/CorpusID:267412288 (cit. alle pp. 1, 12).

[2] Alec Radford et al. «Learning transferable visual models from natural language supervision». In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763 (cit. alle pp. 2, 10).

[3] International Skin Imaging Collaboration. *ISIC Archive - Dermoscopic Images Gallery*. Accessed: 2023-10-05. 2023. URL: https://gallery.isic-archive.com/#!/topWithHeader/onlyHeaderTop/gallery?filter=%5B%22image_type%7Cdermoscopic%22%5D (cit. alle pp. 3, 5, 14).

[4] Veronica Rotemberg et al. «A patient-centric dataset of images and metadata for identifying melanomas using clinical context». In: *Scientific data* 8.1 (2021), p. 34 (cit. a p. 4).

[5] Hadi Zanddizari, Nam Nguyen, Behnam Zeinali e J Morris Chang. «A new preprocessing approach to improve the performance of CNN-based skin lesion classification». In: *Medical & Biological Engineering & Computing* 59.5 (2021), pp. 1123–1131 (cit. a p. 4).

[6] Wei-Hung Weng, Jonathan Deaton, Vivek Natarajan, Gamaleldin F Elsayed e Yuan Liu. «Addressing the real-world class imbalance problem in dermatology». In: *Machine learning for health*. PMLR. 2020, pp. 415–429 (cit. a p. 4).

[7] Hubert Pehamberger, Adi Steiner e Klaus Wolff. «In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions». In: *Journal of the American Academy of Dermatology* 17.4 (1987), pp. 571–583 (cit. a p. 6).

[8] Naheed R Abbasi, Heather M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Issam Osman, Alfred W Kopf e David Polsky. «Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria». In: *Jama* 292.22 (2004), pp. 2771–2776 (cit. a p. 6).

[9] Giuseppe Argenziano, Giuseppe Fabbrocini, Pietro Carli, Vincenzo De Giorgi, Eugenio Sammarco e Massimiliano Delfino. «Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis». In: *Archives of dermatology* 134.12 (1998), pp. 1563–1570 (cit. a p. 6).

[10] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano e Ghassan Hamarneh. «Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets». In: *IEEE Journal of Biomedical and Health Informatics* 23.2 (2019), pp. 538–546. DOI: 10.1109/JBHI.2018.2824327 (cit. a p. 6).

[11] Scott W Menzies, Kerry A Crotty e William H McCarthy. «Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features». In: *Archives of dermatology* 132.10 (1996), pp. 1178–1182 (cit. a p. 6).

[12] John S Henning, Stephen W Dusza, Steven Q Wang, Ashfaq A Marghoob, Harold S Rabinovitz, David Polsky e Alfred W Kopf. «The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy». In: *Journal of the American Academy of Dermatology* 56.1 (2007), pp. 45–52 (cit. a p. 6).

[13] Yinhao Wu, Bin Chen, An Zeng, Dan Pan, Ruixuan Wang e Shen Zhao. «Skin Cancer Classification With Deep Learning: A Systematic Review». In: *Frontiers in Oncology* 12 (2022). ISSN: 2234-943X. DOI: 10.3389/fonc.2022.893972. URL: https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2022.893972 (cit. a p. 7).

[14] Nasser A. AlSadhan, Shatha Ali Alamri, Mohamed Maher Ben Ismail e Ouiem Bchir. «Skin Cancer Recognition Using Unified Deep Convolutional Neural Networks». In: *Cancers* 16.7 (2024). ISSN: 2072-6694. DOI: 10.3390/cancers16071246. URL: https://www.mdpi.com/2072-6694/16/7/1246 (cit. a p. 7).

[15] Flavia Grignaffini, Francesco Barbuto, Lorenzo Piazzo, Maurizio Troiano, Patrizio Simeoni, Fabio Mangini, Giovanni Pellacani, Carmen Cantisani e Fabrizio Frezza. «Machine Learning Approaches for Skin Cancer Classification from Dermoscopic Images: A Systematic Review». In: *Algorithms* 15.11 (2022). ISSN: 1999-4893. DOI: 10.3390/a15110438. URL: https://www.mdpi.com/1999-4893/15/11/438 (cit. a p. 7).

[16] José Ariel Camacho-Gutiérrez, Selene Solorza-Calderón e Josué Álvarez-Borrego. «Multi-class skin lesion classification using prism-and segmentation-based fractal signatures». In: *Expert Systems with Applications* 197 (2022), p. 116671 (cit. a p. 7).

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser e Illia Polosukhin. «Attention is All you Need». In: *Advances in Neural Information Processing Systems*. A cura di I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan e R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. a p. 8).

[18] Alexey Dosovitskiy et al. «An image is worth 16x16 words: Transformers for image recognition at scale». In: *arXiv preprint arXiv:2010.11929* (2020) (cit. a p. 9).

[19] Pinecone. *Vision Transformers: A Comprehensive Guide*. Accessed: 2024-07-16. 2024. URL: https://www.pinecone.io/learn/series/image-search/vision-transformers/ (cit. a p. 9).

[20] Pinecone. *CLIP: Connecting Text and Images with Transformers*. Accessed: 2024-07-16. 2024. URL: https://www.pinecone.io/learn/series/image-search/clip/ (cit. a p. 10).

[21] Corinna Cortes e Vladimir Vapnik. «Support-vector networks». In: *Machine learning* 20.3 (1995), pp. 273–297 (cit. a p. 12).

[22] Pinecone. *Zero-Shot Image Classification with CLIP*. Accessed: 2024-07-16. 2024. URL: https://www.pinecone.io/learn/series/image-search/zero-shot-image-classification-clip/ (cit. a p. 23).

[23] Sirawich Vachmanus, Thanapon Noraset, Waritsara Piyanonpong, Teerapong Rattananukrom e Suppawong Tuarob. «DeepMetaForge: A Deep Vision-Transformer Metadata-Fusion Network for Automatic Skin Lesion Classification». In: *IEEE Access* (2023) (cit. a p. 41).

[24] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner e Alexander Schlaefer. «Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data». In: *MethodsX* 7 (2020), p. 100864 (cit. a p. 41).

[25] Dina Nur Anggraini Ningrum, Sheng-Po Yuan, Woon-Man Kung, Chieh-Chen Wu, I-Shiang Tzeng, Chu-Ya Huang, Jack Yu-Chuan Li e Yao-Chin Wang. «Deep learning classifier with patient's metadata of dermoscopic images in malignant melanoma detection». In: *Journal of multidisciplinary healthcare* (2021), pp. 877–885 (cit. a p. 41).

[26] SP Godlin Jasil e V Ulagamuthalvi. «A hybrid CNN architecture for skin lesion classification using deep learning». In: *Soft Computing* (2023), pp. 1–10 (cit. a p. 41).