



CORSO DI LAUREA IN INFORMATICA

Nevus AI

Autori:

Michele Antonio Annunziata

Vito Vernellati

Docenti:

Prof. Polese Giuseppe

Prof.ssa Caruccio Loredana

Contents

1	Introduzione al problema	4
1.1	Il Machine Learning e la medicina di precisione	4
1.2	La diagnosi precoce dei melanomi	4
1.3	Il problema degli approcci IA in ambito medico	4
1.4	Il problema affrontato	5
1.5	Il problema dell’explainability	5
1.5.1	In ambito clinico	5
1.6	Il problema dei dati mancanti e sbilanciati	5
2	Data understanding and data preparation	6
2.1	Panoramica delle opzioni disponibili	6
2.1.1	Approccio con sole immagini	6
2.1.2	Approccio con immagini e feature testuali	6
2.1.3	Approccio scelto e motivazioni	6
2.2	Data gathering	7
2.2.1	Dataset construction	7
2.2.2	Analisi dei dataset costruiti	8
2.3	Data preparation	9
2.3.1	Pulizia dei dati	9
2.3.2	Feature scaling	10
2.3.3	Bilanciamento delle classi	10
3	Sviluppo del modello	11
3.1	Quale modello scegliere	11
3.1.1	Due differenti approcci	11
3.2	Come valutare	12
3.3	Implementazione	12
4	Training e evaluation	14
5	Deployment, usability e explainability	15

List of Figures

1	Bilanciamento delle classi iniziale dei dataset	8
2	Bilanciamento delle classi iniziale dei dataset	11

List of Tables

1	Metadati utilizzati per lo splitting	8
---	--	---

1 Introduzione al problema

1.1 Il Machine Learning e la medicina di precisione

Il tema della medicina di precisione con l'ausilio dell'intelligenza artificiale è una tematica sempre più trattata dalla comunità scientifica. L'obiettivo principale di molti studi è l'applicazione di metodi di Machine Learning per definire modelli computazionali in grado di supportare diagnosi cliniche e migliorare la prevenzione. In particolare, l'analisi di immagini mediche attraverso modelli "*intelligenti*" rappresenta una svolta significativa nel settore, consentendo di automatizzare processi diagnostici.

1.2 La diagnosi precoce dei melanomi

Nel contesto della dermatologia, il melanoma cutaneo rappresenta una delle sfide più critiche poiché la sua aggressività può risultare fatale se non diagnosticato precocemente. Seppur una diagnosi precoce aumenterebbe significativamente le possibilità di sopravvivenza, questa richiede competenze specialistiche e strumenti adeguati che non sempre sono disponibili in tutti i contesti clinici.

L'utilizzo di applicazioni che supportano tecnologie di Machine Learning rappresenta un supporto immediato, concreto e prezioso per migliorare l'identificazione delle lesioni cutanee sospette. Questi modelli permettono anche agli esperti di dominio che operano in contesti con risorse limitate, di accedere ad una diagnosi precoce con cui si distingue lesioni benigne e lesioni maligne.

Tutto ciò contribuisce alla riduzione di mortalità associata al melanoma e, più in generale, al miglioramento della qualità complessiva delle cure mediche.

1.3 Il problema degli approcci IA in ambito medico

Come per qualsiasi approccio di Machine Learning e Intelligenza Artificiale, l'applicazione in ambito medico presenta sfide specifiche.

Una delle principali difficoltà è la reperibilità di dati di alta **qualità**: le immagini di lesioni cutanee devono essere ben annotate e rappresentative della popolazione target per consentire un elevato livello di generalizzazione del modello. La dicotomia tra i dati utilizzati dal mondo della "*ricerca*" e i dati utilizzati dal mondo della "*produzione*" rende molti modelli inutilizzabili nell'effettivo, soprattutto a causa dell'eccessivo apporto di dati di sintesi.

Un altro aspetto cruciale è senza dubbio l'**explainability**: bisogna cercare di fornire agli utilizzatori del modello una qualche motivazione per cui l'algoritmo di Machine Learning ha fornito una certa

predizione. Ciò è fondamentale per conquistare la fiducia, e un certo grado di confidenza, degli utenti (medici e pazienti nella fattispecie).

1.4 Il problema affrontato

Il progetto **Nevus AI** mira a sviluppare un modello di Machine Learning in grado di classificare immagini di nei in due categorie principali: benigni e maligni. Partendo da un dataset composto da immagini dermatologiche ben annotate, l'obiettivo è creare un modello accurato, robusto e interpretabile, ponendo particolare enfasi sull'analisi dei dati e sulla trasparenza/comprendibilità.

Il progetto è corredato dalla repository disponibile qui.

1.5 Il problema dell'explainability

Come già accennato precedentemente, uno dei problemi principali che rende inutilizzabili nella realtà dei modelli ottimi in letteratura, è l'explainability, la quale misura quanto il processo decisionale su cui si basa il modello sia comprensibile a chi lo utilizza. Maggior explainability comporta maggior fiducia, maggior comprensione, maggior benefici in ambito legale e maggior capacità di rilevare eventuali bias.

1.5.1 In ambito clinico

Un dermatologo che non riesce a comprendere perché un modello ha classificato una lesione come maligna difficilmente farà affidamento su di esso. Analogamente, un paziente potrebbe non accettare una diagnosi basata esclusivamente su una "scatola nera" senza una spiegazione convincente. Da ciò si evince quanto sia importante soffermarsi ed evidenziare il grado di explainability.

1.6 Il problema dei dati mancanti e sbilanciati

Un altro problema rilevante è sicuramente la disponibilità di dati di qualità. In ambito dermatologico, le immagini possono essere fortemente sbilanciate tra classi (ad esempio, potrebbero esserci più campioni di nei benigni ai maligni) o contenere informazioni incomplete. Questo porta inevitabilmente il modello a ridurre le sue capacità predittive. Naturalmente, la progettazione di modelli di Machine Learning capaci di generalizzare richiede una gestione rigorosa ed accurata dei dati iniziali.

2 Data understanding and data preparation

2.1 Panoramica delle opzioni disponibili

Durante la fase di analisi iniziale, sono stati identificati e considerati due approcci principali per sfruttare i dati raccolti nel contesto del problema.

2.1.1 Approccio con sole immagini

Questo approccio si basa esclusivamente sull'utilizzo delle immagini delle lesioni cutanee come input per il modello di classificazione. I dati visivi sono particolarmente indicativi per il riconoscimento delle caratteristiche morfologiche e dei pattern associati alle lesioni cutanee.

L'impiego di strumenti sofisticati di Machine Learning come le reti neurali convoluzionali rende questo approccio particolarmente efficace, poiché le immagini forniscono una base ricca di informazioni che possono essere sfruttate allo scopo di estrarre *automaticamente* le caratteristiche maggiormente rilevanti. L'esclusione di eventuali dati testuali tuttavia, potrebbe limitare la capacità di integrazione di aspetti contestuali e complementari che potrebbero arricchire l'analisi.

2.1.2 Approccio con immagini e feature testuali

In questo approccio, oltre alle immagini, vengono utilizzate feature testuali come input aggiuntivi per il modello. Tali informazioni possono includere generiche annotazioni descrittive che servono come supporto al miglioramento delle capacità predittive del modello.

Tuttavia, l'uso di dati testuali può presentare problematiche significative, quali incompletezza, incoerenza o mancanza di standardizzazione, le quali possono inevitabilmente distorcere il contenuto informativo introducendo rumore e compromettendo l'efficacia dei risultati ottenuti.

2.1.3 Approccio scelto e motivazioni

Dopo un'attenta valutazione delle opzioni disponibili ed un'analisi costo-benefici, è stato deciso di adottare l'**approccio con sole immagini**.

La scelta è stata motivata principalmente dai seguenti fattori:

- **qualità delle immagini:** le immagini presenti nelle varie collezioni sono di alta qualità e rappresentano un punto di partenza solido per l'analisi visiva approfondita delle lesioni cutanee.

Concentrare gli sforzi computazionali sull'elaborazione di immagini consente una focalizzazione sui pattern visivi, i quali rappresentano il principale elemento diagnostico.

- **ottimizzazione delle risorse e dei tempi:** la pipeline basata su sole immagini risulta essere maggiormente semplice e scalabile rispetto ad una pipeline multimodale. Ciò consente di rispettare i vincoli di tempo e risorse stabiliti in fase di pianificazione.
- **limitazioni dei dati testuali:** Le informazioni testuali disponibili sono incomplete e non sempre standardizzate, il che potrebbe introdurre ambiguità o rumore nei dati. Questa mancanza di coerenza ha influenzato la scelta di evitare l'integrazione di dati testuali.

2.2 Data gathering

Il dataset utilizzato e le risorse necessarie per proseguire con il completamento del progetto sono stati raccolti dalla piattaforma **ISIC Archive** disponibile [qui](#).

In particolare sono stati inizialmente esaminati ed esplorati i dataset presenti nelle collezioni intitolate "**BCN20000**" (disponibile [qui](#)) e "**HAM10000**" (disponibile [qui](#)).

Entrambi i dataset includono due tipi di dati:

- immagini delle lesioni cutanee. Tutta la raccolta *HAM10000* presenta immagini di dimensioni 600×450 , mentre quelle nel dataset *BCN20000* sono di dimensioni 1024×1024 ;
- metadati testuali che forniscono informazioni aggiuntive sulle caratteristiche cliniche e contestuali delle lesioni.

Nelle sezioni successive saranno riportate tutte le scelte adottate con le relative motivazioni.

2.2.1 Dataset construction

Per rispettare l'approccio scelto, descritto nelle sezioni precedenti, è stato deciso di costruire internamente un dataset contenente esclusivamente immagini di lesioni cutanee, selezionate dai dataset *HAM10000* e *BCN20000*.

La costruzione dei dataset è stata effettuata tramite uno script python sviluppato internamente. Lo script ha eseguito le seguenti operazioni:

- **lettura dei metadati** per ciascuno dei due dataset. In Tabella 1 sono riportate le feature esaminate per lo splitting.

isic_id	benign_malignant	diagnosis_1	diagnosis_2	diagnosis_3
---------	------------------	-------------	-------------	-------------

Table 1: Metadati utilizzati per lo splitting

- **creazione delle cartelle annotate:** per ciascun dataset, sono stati creati tre gruppi principali di immagini basati sulle etichette fornite.

1. **nei_benigni:** contenente le immagini classificate come benigni.
2. **nei_maligni:** contenente le immagini classificate come maligni.
3. **nei_indeterminati:** contenente le immagini classificate come indeterminate.

Il risultato finale ottenuto è la creazione di due cartelle distinte contenente le immagini presenti nei dataset "*HAM10000*" e "*BNC20000*".

2.2.2 Analisi dei dataset costruiti

Dopo la costruzione dei dataset mediante lo script che ha suddiviso le immagini in base alle etichette presenti nei metadati, è stata effettuata un'analisi dettagliata della distribuzione delle classi in ciascun dataset. I risultati sono riassunti in Figura 1.

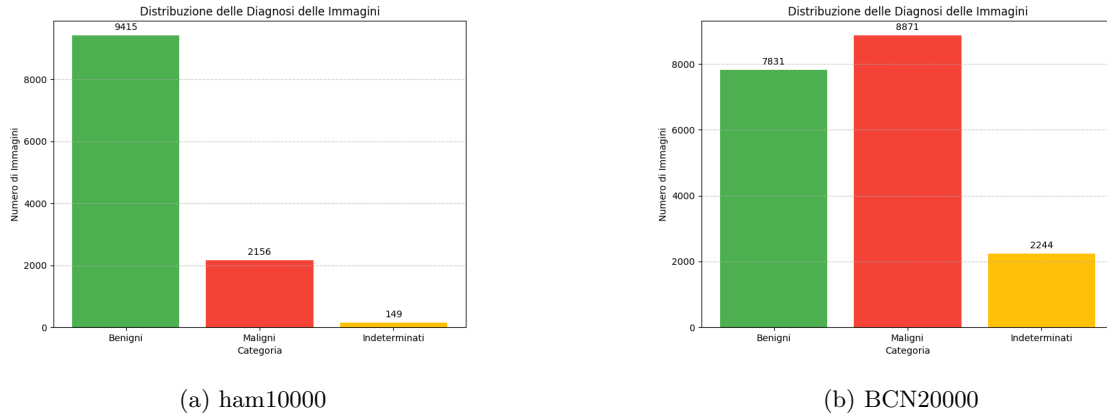


Figure 1: Bilanciamento delle classi iniziale dei dataset

L'analisi della distribuzione ha evidenziato una significativa differenza nel bilanciamento delle classi tra i due dataset. Nel dataset *HAM10000*, la classe dei nei benigni rappresenta la maggioranza, con una proporzione di circa il 80.5% rispetto al 18.5% dei nei maligni. Inoltre circa l'1% è rappresentato da immagini con etichetta "*indeterminato*". Questa forte discrepanza potrebbe influire negativamente

sull'addestramento del modello, portandolo a favorire previsioni sulla classe dominante, con conseguenti problemi di generalizzazione.

Al contrario, nel dataset *BCN20000*, la distribuzione delle classi è più bilanciata, con una porzione circa del 41.5% di nei benigni e del 47% di nei maligni. Inoltre circa l'11.5% è rappresentato da immagini con etichetta "*indeterminato*". Questo bilanciamento offre un punto di partenza più solido per l'addestramento del modello, riducendo la necessità di applicare tecniche di riequilibrio e riducendo il divario tra classe maggioritario e minoritaria.

Il problema delle classi non bilanciate. In generale, quando non si hanno abbastanza campioni per discriminare delle classi differenti, si è in una situazione sgradevole poiché, così facendo, le capacità di generalizzare del modello vanno a decrescere per le classi minoritarie. Una situazione ideale sarebbe quella in cui tra le classi ci sia un totale bilanciamento in termini di numero di campioni.

Scelta effettuata. Sulla base delle analisi eseguite, è stato scelto di utilizzare il dataset costituito dalle immagini presenti in *BCN20000*. Sono riportate di seguito le motivazioni principali.

- **Bilanciamento delle classi.** La distribuzione delle immagini nei benigni e maligni nel dataset *BCN20000* è maggiormente equilibrata rispetto al dataset *HAM10000*, è ciò consente di ridurre notevolmente il rischio di bias verso la classe dominante.
- **Rappresentatività dei dati maligni.** La maggiore presenza di immagini di nei maligni nel dataset costruito a partire da *BCN20000* consente al modello di apprendere più efficacemente le caratteristiche distintive di queste lesioni, migliorando la sensibilità diagnostica.

Questa decisione mira a garantire che il modello possa esser robusto ed il più capace possibile a discernere le classi.

2.3 Data preparation

2.3.1 Pulizia dei dati

Pulire i dati per renderli utilizzabili da un modello di Machine Learning è un passo fondamentale per garantire qualità e affidabilità. Pulizia dei dati significa capire in che modo sopravvivere alla mancanza o al rumore dei dati di partenza.

Nel contesto di questo progetto, sono state adottate le seguenti misure:

- **Esclusione delle immagini indeterminate.** Tutte le immagini presenti nella cartella denominata "nei_indeterminati" sono state scartate. Questo perché le etichette "indeterminate" non forniscono informazioni utili per l'addestramento del modello e potrebbero introdurre ambiguità durante la fase di apprendimento.
- **Verifica delle immagini corrotte.** Le immagini non leggibili o corrotte sono state eliminate per evitare errori durante il caricamento e la fase di pre-elaborazione (tale operazione è stata effettuata tramite lo script python per lo splitting delle cartelle citato nelle sezioni precedenti)

2.3.2 Feature scaling

Generalmente, siccome i domini delle caratteristiche sono differenti, un passo fondamentale è quello di modificare i valori in scala per evitare che le scale più ampie possano monopolizzare la fase di apprendimento. Attributi con grosse differenze di scala provocano problemi per il modello, il quale potrebbe sottostimare/sovrastimare la rilevanza di una caratteristica.

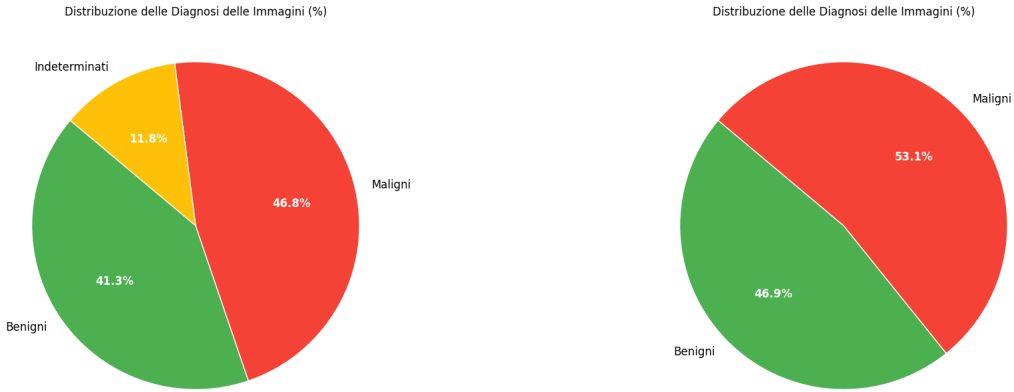
Nel caso di Nevus AI. Non essendo presenti feature numeriche tradizionali, la fase di feature scaling è stata applicata esclusivamente alle immagini, normalizzando i valori dei pixel. I valori sono stati scalati nell'intervallo $[0, 1]$ dividendo ciascun pixel per 255.

Questo processo migliora la stabilità numerica durante l'addestramento del modello e rende il dataset coerente per le tecniche di deep learning.

2.3.3 Bilanciamento delle classi

Come già detto nelle sezioni precedenti, lo squilibrio tra le classi è un problema comune nei dataset. In Figura 2 è mostrata la distribuzione delle classi del dataset ricavato dalle immagini in "BCN20000".

Per affrontare questo problema e migliorare le capacità di generalizzazione del modello, si è deciso di aumentare i campioni della classe minoritaria, adottando l'algoritmo SMOTE (*Synthetic Minority Oversampling Technique*), il quale consente di partire dalle immagini originali della classe "benigni" e generare sinteticamente nuove immagini, ad esempio tramite trasformazioni geometriche.



(a) BCN20000 considerando "indeterminati"

(b) BCN20000 escludendo "indeterminati"

Figure 2: Bilanciamento delle classi iniziale dei dataset

3 Sviluppo del modello

3.1 Quale modello scegliere

Il task in esame, la classificazione delle lesioni cutanee, richiede un approccio robusto e accurato per distinguere tra lesioni benigne e maligne. Per tale task è stato deciso di adottare una tipologia di architettura allo stato dell'arte per lavorare con le immagini, ossia le CNN (*Convolutional Neural Network*).

3.1.1 Due differenti approcci

Considerando l'efficacia delle reti neurali convoluzionali (CNN) nell'elaborazione delle immagini, sono state esplorate due opzioni principali:

- **starting from scratch**, un approccio che comporta la costruzione da zero di una CNN, progettata specificamente per il dataset di lesioni cutanee. Richiede un numero significativo di dati e tempo di addestramento per ottenere risultati competitivi;
- **sequential transfer learning**, una tecnica in cui un modello pre-addestrato in un dominio D_s su un task T_s viene riutilizzato, adattato o perfezionato per risolvere un nuovo task T_t in un dominio D_t . Questo metodo permette di risparmiare tempo e risorse, e di trasferire le

conoscenze acquisite da task generici di classificazione delle immagini, con risultati spesso più rapidi e accurati.

Approccio scelto. Dopo un'analisi comparativa, è stato deciso di adottare l'approccio di sequential transfer learning. Questa scelta è motivata dall'alta qualità delle reti preaddestrate disponibili e dalla loro comprovata efficacia in contesti simili.

Fasi principali. Il sequential transfer learning si basa su un adattamento progressivo da un task sorgente a un task target, seguendo diverse fasi fondamentali:

- 1) Il modello viene addestrato su un dataset sorgente *grande e generico*, così da formare la base di ciò che si deve apprendere.
- 2) Successivamente il modello pre-addestrato viene ulteriormente raffinato utilizzando un dataset target più specifico.
 - 2.1) **Congelamento** dei layer. I primi strati della rete che costituiscono le caratteristiche generali vengono congelati per preservare le conoscenze già acquisite.
 - 2.2) **Fine-tuning**. Gli strati finali vengono ulteriormente addestrati con i dati target per apprendere caratteristiche più specifiche. Durante questa fase i pesi di questi strati vengono aggiornati per affinare il modello e migliorare la precisione per il compito specifico.

3.2 Come valutare

La valutazione dei modelli di classificazione è cruciale per garantire la loro affidabilità e precisione. È stato scelto di utilizzare una **Stratified K-Fold Cross Validation**, che permette di mantenere il bilanciamento delle classi in ogni fold di validazione, migliorando così la robustezza delle stime.

Le metriche di valutazione includeranno l'**accuracy**, la **precision**, la **recall**, la metrica **F1-Score** e la **confusion matrix**.

Ogni fold di validazione genererà queste metriche, consentendo un'analisi approfondita delle performance del modello.

3.3 Implementazione

Per l'implementazione del modello e delle metriche di valutazione, sono state utilizzate le seguenti librerie:

- **TensorFlow/Keras**, per la creazione, addestramento e valutazione delle CNN.
- **scikit-learn**, per il pre-processing dei dati e la gestione della cross-validation.
- **pandas** e **numpy**, per il caricamento e la manipolazione dei dataset.
- **matplotlib** e **seaborn**, per la creazione di grafici e visualizzazioni delle metriche.
- **imblearn**, per applicare tecniche di oversampling.

4 Training e evaluation

5 Deployment, usability e explainability