POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Simulazione di Dati Radiomici per una Pipeline Robusta e Validazione dei Metodi Usati

TESI DI LAUREA TRIENNALE IN
INGEGNERIA BIOMEDICA

Autore:
Gabriella Bertolino

Student IDs: 10736705
Advisor: Chiara Paganelli
Tutor: Mariagrazia Monteleone, Lorenzo Cederle
Academic Year: 2023-24

Summary

The project falls within the field of radiomics, a discipline that concerns the extraction and analysis of quantitative features from medical images providing minable data not detectable by simple visual inspection. To achieve this goal, specific characteristics (called features) are extracted from the images and, through the use of sophisticated algorithms, can be used to achieve several clinical tasks (e.g., screening, prognosis or diagnosis). However, for these algorithms to be properly trained and work effectively a large amount of data is required, which is often difficult to obtain.

Therefore, this work aims to create synthetic tabular datasets of radiomic features, using two distinct methods and conducting a comparative analysis between them.

To ensure the quality of the simulated data, it is essential that they faithfully reflect the features of the real dataset. In the context of our project, data were retrieved and transformed into tabular format from CT images obtained from the NSCLC (Non-Small Cell Lung Cancer) dataset of the Cancer Imaging Archive. The actual dataset consists of a series of observations, each associated with features that correlate with the target variable, represented by lung cancer histology. Accordingly, the dataset can be divided into two subsamples based on diagnosis: the first includes Adenocarcinoma histology (ADN), while the second includes Squamous Cell Carcinoma (CCS) or not otherwise specified (NOS).

After a literature review on the most used methods for generating synthetic tabular data in the context of radiomics, we decided to employ a traditional method, namely the Monte Carlo algorithm implemented by Vale and Maurelli, which was compared with a more innovative method based on Deep Learning, known as CTGAN (Conditional Tabular Generative Adversarial Network). After carefully evaluating the quality of the data generated, we finally highlighted both the limitations and advantages of the two methods used, thus providing a comprehensive overview of their respective efficiency and applicability.

Key-words: radiomics, synthetic data generation.

Sommario

Il progetto si colloca nel campo della radiomica, una disciplina che consiste nell'analisi delle immagini mediche finalizzata all'estrazione di informazioni quantitative, informazioni che non sono rilevabili tramite la semplice osservazione visiva da parte dell'operatore. Per raggiungere questo obiettivo, vengono estratte dalle immagini specifiche caratteristiche (dette features) e, mediante l'uso di sofisticati algoritmi, è possibile realizzare diversi task clinici (ad esempio screening, prognosi o diagnosi). Tuttavia, affinché questi algoritmi possano essere adeguatamente addestrati e funzionino in modo efficace, è necessaria una grande quantità di dati, che spesso risulta difficile da reperire.

Questo lavoro si propone pertanto di creare dataset sintetici tabulari contenenti features radiomiche impiegando due metodi distinti, per poi confrontarli e validarli.

Per garantire la qualità dei dati simulati, è fondamentale che essi riflettano fedelmente le caratteristiche del dataset reale. Nel contesto del nostro progetto, i dati sono stati recuperati e trasformati in formato tabulare partendo da immagini TAC ottenute dal dataset NSCLC (Non-Small Cell Lung Cancer) del Cancer Imaging Archive. Il dataset reale è composto da una serie di osservazioni, ciascuna delle quali è associata a features che presentano una correlazione con la variabile target, rappresentata dall'istologia del tumore al polmone. Di conseguenza, il dataset può essere suddiviso in due sottocampioni in base alla diagnosi: il primo comprende l'istologia di Adenocarcinoma (ADN), mentre il secondo include l'istologia di Carcinoma a Cellule Squamose (CCS) o istologia non altrimenti specificata (NOS).

Dopo aver condotto una ricerca bibliografica sui metodi più utilizzati per la generazione di dati tabulari sintetici nel contesto della radiomica, abbiamo deciso di impiegare un metodo tradizionale, ovvero l'algoritmo di Monte Carlo implementato da Vale e Maurelli, che è stato confrontato con un metodo più innovativo basato sul Deep Learning, il CTGAN (Conditional Tabular Generative Adversarial Network).

Dopo aver attentamente valutato la qualità dei dati generati, abbiamo infine evidenziato sia le limitazioni che i vantaggi dei due metodi utilizzati, fornendo così una panoramica completa della loro rispettiva efficienza e applicabilità.

Parole chiave: radiomica, generazione di dati sintetici

Contenuti

Introduzione

Negli ultimi anni la disciplina della radiomica nel contesto dell'oncologia ha acquisito sempre più importanza. La radiomica è un campo di ricerca in rapida evoluzione che si occupa dell'estrazione di metriche quantitative, le cosiddette features radiomiche, da immagini mediche. Le feature radiomiche catturano le caratteristiche dei tessuti e della lesione, come l'eterogeneità e la forma, e possono essere utilizzate, da sole o in combinazione con dati demografici, istologici, genomici o proteomici, per la risoluzione di problemi clinici.[1] Queste informazioni quantitative estratte dalle immagini mediche sono quindi date in input ad algoritmi di Machine Learning (ML) che assisteranno, con i loro risultati, lo svolgimento di task clinici essenziali, come screening, prognosi, diagnosi, pianificazione del trattamento e valutazione della risposta al trattamento stesso e giocando quindi un grande ruolo all'interno della medicina personalizzata. [2]

Il workflow di radiomica si articola in diversi passaggi: acquisizione dell'immagine, segmentazione della regione d'interesse, estrazione delle features e successiva analisi. La prima fase prevede l'acquisizione di immagini. Da questa immagine viene definito una regione di interesse (ROI) utilizzando tecniche di segmentazione manuale o semiautomatica. Successivamente, vengono estratte le caratteristiche quantitative dell'immagine dalla regione precedentemente definita che vengono poi adeguatamente analizzate e processate.[3]

Sebbene la radiomica abbia dimostrato il suo potenziale in numerosi studi, la sua trasposizione nella pratica clinica è stata piuttosto lenta e le ragioni sono molteplici. [4] Infatti una sfida che la radiomica deve ancora affrontare è la mancanza di un'adeguata validazione esterna dei modelli di ML. Come delineato da Jacobs et al in Opportunities and Challenges of Synthetic Data Generation in Oncology, l'imaging del cancro è l'area in cui sono stati compiuti i maggiori progressi, sebbene esistano ancora sfide importanti, tra cui le dimensioni ridotte e la complessità delle lesioni tumorali, la variabilità tra gli osservatori, l'eterogeneità intra e inter-tumorale, la difficoltà di annotare, delineare ed etichettare gli studi di imaging oncologico su larga scala, lo sbilanciamento dei dati e la difficoltà di raccogliere grandi insiemi di

dati previo consenso[5]. Spesso dunque, variazioni tra macchinari utilizzati o diversità nella qualità delle immagini portano ad una grande eterogeneità dei dati, la laboriosità del processo di annotazione da parte di esperti porta ad una disponibilità limitata dei dati e la sensibilità dei dati stessi comporta problemi dovuti all'etica e alla privacy; tutte caratteristiche che rendono l'ottenimento di dati medici particolarmente difficoltoso. Dall'altra parte invece gli algoritmi di ML per performare in modo adeguato e fornire risultati coerenti hanno bisogno di ingenti quantità di dati per essere addestrati.

Per ovviare a questa mancanza di dati, una via che si può intraprendere è l'utilizzo di dati sintetici (SD), ovvero dati artificiali generati da un modello addestrato (o costruito) per replicare dati reali e imitando quindi le caratteristiche matematiche e statistiche dei dati di input. Dal momento che esistono diversi metodi per generare insiemi di SD, è importante capire in che modo questi metodi di generazione dei SD differiscono l'uno dall'altro e quale sia il più appropriato in base al contesto di applicazione.

Il primo passo nella generazione di set di SD è la comprensione della natura della verità di base, cioè da dove derivano i dati di riferimento utilizzati. Pertanto, i metodi di generazione sono generalmente classificati in tre categorie in base all'identificazione della verità di base: [6]

- Metodi guidati dalla conoscenza/knowledge-driven (basati sulla conoscenza derivata da documenti accademici e dall'esperienza umana)
- Metodi guidati dai dati/data-driven(basati su dati reali)
- Metodi ibridi/hybrid (la verità di base è un amalgama di teoria e dati reali)

La nostra ricerca si concentra sulla generazione di insiemi di dati tabulari a partire da dati reali. Dunque, una volta identificata la verità di base e il tipo di metodo generativo più adatto, si può procedere alla generazione vera e propria del set di SD.

Nel momento in cui si genera un dataset a partire da dati reali, una delle prime sfide che si può incontrare è lo sbilanciamento tra le diverse classi presenti all'interno del dataset, vi possono essere infatti dati sottorappresentati e tale evento risulta essere un problema in particolare nel momento in cui si utilizzano per la generazione dei metodi basati su ML, che come sarà poi spiegato successivamente, risultano essere i metodi più utilizzati oggi per la generazione di SD. Questa sottorappresentazione

può causare mancanza di diversità, bias nei modelli utilizzati che prediligono i dati maggioritari e rischio di generare dati non realistici.

Per evitare che l'algoritmo favorisca una certa classe possono essere utilizzati dei metodi per riequilibrare il campione, come undersampling e oversampling. Nel caso dell'undersampling si ridimensiona una classe estraendo da essa un suo sottoinsieme, invece, per l'oversampling si cercano di generare dati simili in modo da bilanciare le classi. Uno dei metodi più utilizzati con questa ultima tecnica è il metodo SMOTE (Synthetic Minority Over-sampling Technique). Il vantaggio che ha l'oversampling rispetto all'undersampling sta nel fatto che non viene tolta nessuna informazione dei dati veri per poter bilanciare le classi, ma genera nuovi esempi sintetici della classe minoritaria anziché semplicemente replicare quelli esistenti.[7]

Effettuando una ricerca in letteratura è emerso che è possibile suddividere i metodi di generazione di dati fittizi in due categorie principali: metodi tradizionali e metodi di Deep Learning (DL). I metodi tradizionali di generazione di SD racchiudono sia metodi statistici che di apprendimento automatico: nei metodi statistici, i dati originali vengono studiati con lo scopo di individuare la loro distribuzione statistica per generare dati fittizi con la medesima distribuzione. Nei modelli di apprendimento automatico, invece, si studiano i pattern che poi si ritroveranno nei dati fittizi.[6] [7] [8]

I principali metodi tradizionali sono:

Monte Carlo

La simulazione Monte Carlo è una tecnica computazionale che utilizza il campionamento casuale per modellare e analizzare sistemi o processi complessi, fornendo stime dei possibili risultati e delle loro probabilità. Questo metodo verrà approfondito nel capitolo successivo.

Categorical Latent Gaussian process

La tecnica del Processo Gaussiano Latente Categorico (CLGP) è un metodo che unisce concetti di processi gaussiani latenti (GP) con variabili categoriali per analizzare e generare dati che non sono esclusivamente numerici ma possono avere valori discreti o essere categorizzati.

Multiple Imputation

Questo metodo si concentra su insiemi di dati che contengono informazioni sensibili. Per poter lavorare con essi, infatti, i dati sensibili vengono caratterizzati come mancanti, per poi essere sostituiti con dati fittizi che rispecchino il più possibile quelli originali.

In secondo luogo vi sono i metodi innovativi o di DL. I metodi di DL sono basati su reti neurali artificiali utilizzate per apprendere e analizzare quantità ingenti di dati. Questi metodi sono diventati sempre più popolari, in quanto sono in grado di catturare e riprodurre la complessità dei dati reali, consentendo di generare campioni che sono difficili da distinguere dai dati reali. [9] [10]

In particolare, per la generazione di SD i metodi più utilizzati sono:

Generative Adversarial Networks (GAN)

Consistono principalmente in due reti neurali (generatore e discriminatore), che imparano a generare SD di alta qualità attraverso un processo di addestramento avversario e competitivo. Questo metodo verrà approfondito nel capitolo successivo. [11]

Variational AutoEncoders (VAE)

I Variational Autoencoders (VAE) combinano elementi di autoencoder con tecniche di inferenza bayesiana per apprendere e generare dati che assomigliano ai dati di addestramento.

Reti Bayesiane

Utilizzano algoritmi basati sull'approccio probabilistico di Bayes per generare campioni di dati che rispecchiano le distribuzioni probabilistiche dei dati reali.

In generale, il metodo utilizzato per la generazione di SD dipende sia dallo scopo del lavoro e sia i metodi tradizionali che quelli più innovativi presentano vantaggi e svantaggi. Infatti, i primi presentano una struttura matematica semplice, sono più

facili da comprendere e implementare e sono inoltre computazionalmente poco costosi e quindi possono essere eseguiti anche su hardware meno potenti. D'altro canto i metodi tradizionali sono più dispendiosi in termini di tempo (rispetto a metodi di DL che sono progettati per lavorare in parallelo) e possono avere difficoltà a catturare la complessità nei dati reali, cioè mancano della flessibilità necessaria per adattarsi a dataset complessi e variabili. I metodi di DL hanno una buona scalabilità, sono estremamente potenti nel catturare strutture complesse e relazioni non lineari nei dati e in generale, un buon rapporto costo-efficacia; tuttavia richiedono anche processori appositamente progettati e risorse computazionali significative.[7]

L'uso dei SD in ambito clinico presenta opportunità interessanti, come il costo contenuto, la possibilità di produrre istanze specifiche e di avere un accesso immediato ai dati, solo per citarne alcune. Gli ambiti clinici interessati sono molteplici: la ricerca medica, ad esempio, per malattie rare per cui i dati reali potrebbero scarseggiare; la simulazione di trial clinici, in modo da aiutare a prevedere i risultati diagnostici; o ancora la formazione del personale medico, cioè utilizzare SD per creare scenari di formazione realistici per medici e infermieri.

Tuttavia, gli SD pongono anche sfide importanti, come la necessità della valutazione della qualità e del realismo; infatti più alto è il realismo dei SD, meglio possono sostituire gli originali nelle applicazioni.

L’altra area di ricerca attiva è la valutazione della qualità dei SD e le metriche proposte possono essere classificate in tre caratteristiche principali: fedeltà, diversità e generalizzazione.

La fedeltà misura il grado di realismo degli SD e si occupa della consistenza dei dati. Questo si riferisce alla misura in cui questi dati mantengono proprietà e relazioni logiche, statistiche e strutturali simili a quelle dei dati reali da cui sono derivati o che cercano di emulare. In generale, le due opzioni più popolari per misurare la fedeltà sono i metodi computazionali e umani. Dal punto di vista computazionale, si utilizzano metodi statistici e intervalli di confidenza per confrontare la distribuzione dei dati reali e la distribuzione dei SD. È anche una pratica comune valutare la fedeltà dei dati chiedendo agli esperti di decidere se un campione sia reale o sintetico e di riferire la loro frequenza di successo. Tuttavia troppa fedeltà può portare all'overfitting, dove i SD replicano non solo i pattern generali ma anche il rumore

specifico dei dati di addestramento.

La misurazione della diversità è fondamentale per valutare la varietà all'interno dei SD, in quanto garantisce che tutti i sottogruppi dei dati originali siano rappresentati in modo appropriato, il che significa che i SD riflettono accuratamente i dati originali, consentendo ai ricercatori di trarre conclusioni significative. I SD devono quindi includere casi rari e limiti estremi per testare i modelli in condizioni meno comuni, ma pur sempre possibili. Ma, anche in questo caso, un'eccessiva diversità può impattare negativamente sulla fedeltà, se i dati diventano troppo variegati e non rappresentativi dei dati reali.

Infine, la generalizzazione può essere utilizzata per valutare l'autenticità dei SD e dunque la capacità dei SD di rappresentare pattern che si estendono oltre il dataset di addestramento. Se i SD sono troppo simili ai dati reali, possono compromettere la privacy dei pazienti. D'altra parte, un'eccessiva generalizzazione nel processo di SD potrebbe non tenere conto di gruppi sottorappresentati nei dati reali, e quindi richiederebbe sacrifici in termini di fedeltà e diversità. [4]

Si evince che una generazione di SD efficace richiede un attento equilibrio tra fedeltà, diversità e generalizzazione. Ogni aspetto deve essere considerato in relazione agli altri per creare SD utili e affidabili per l'applicazione desiderata.

Quindi, dopo aver generato i SD e prima di sfruttarli, un passo importante da fare è assicurarsi che tutte le sfide che abbiamo elencato in precedenza siano superate dai SD che vogliamo usare, il che significa che superino il test di valutazione della qualità.

Nonostante le sfide menzionate precedentemente, nella letteratura vi sono alcuni risultati promettenti ottenuti con i nuovi metodi di DL e si nota una migliore implementazione all'interno dei metodi statistici.

In particolare, l'ambito medico in cui si sono ottenuti i maggiori progressi è quello dell'imaging dei carcinomi, ovvero del rilevare e caratterizzare i tumori con il fine ultimo di migliorare la specificità degli strumenti diagnostici.[4] Questo potrebbe suggerire che la generazione di SD per scopi medici potrebbe essere il futuro della diagnosi e della prognosi medica. Tuttavia, per creare un modello robusto per la validazione delle pipeline, sarà fondamentale capire come i diversi parametri tra i

modelli (statistici o DL), possano influenzare le prestazioni della generazione di SD. Inoltre, si dovrà analizzare quale tipo di caratteristiche sono rilevanti e come sono correlate tra loro. Perciò, verranno generati più dataset sintetici utilizzando diversi metodi e verranno messi a confronto i risultati, per poi stabilire quale sia il più adeguato per il campo della radiomica. Da una parte, il metodo di Vale e Maurelli (VM) [12] per la generazione di dati multivariati non normali è stato comunemente utilizzato per studi radiomici e sarà scelto come metodo di rappresentanza tra quelli tradizionali statistici. Esso è progettato per generare dati che rispettano una specifica matrice di correlazione. Questo è cruciale in radiomica, dove le caratteristiche estratte dalle immagini mediche spesso hanno correlazioni complesse e significative. Mantenere queste correlazioni nei SD assicura realismo dei dati e validità dei risultati. Dall'altra i modelli GAN rappresentano lo stato dell'arte nella generazione di SD e per questo verrà selezionato tra i metodi DL. Con adattamenti come il CTGAN, si possono generare SD tabulari; questi spesso contengono variabili categoriche e numeriche con distribuzioni complesse e relazioni non lineari, che il CTGAN è in grado di catturare. [13]

Lo scopo del nostro lavoro è quindi quello di analizzare un un dataset di features radiomiche estratte da immagini TAC, pre-elaborarlo e quindi utilizzarlo per generare SD attraverso dapprima l'utilizzo del metodo tradizionale VM e poi con il metodo basato su DL CTGAN.

1 Materiali e Metodi

1.1. Il Dataset

Il dataset di features radiomiche utilizzato per questo studio è stato acquisito da una raccolta di immagini TAC disponibili sulla repository open access The Cancer Imaging Archive. In particolare le immagini selezionate si riferiscono al Radiogenomics dataset [14], che include 211 pazienti affetti da *Non-Small-Cell Lung Cancer* (NSCLC), fornendoci, quindi, una base solida per le nostre analisi.

Il dataset specifico che è stato utilizzato è un sottocampione del dataset originale, esso è composto da 144 osservazioni, tra le immagini TAC dei 211 pazienti infatti sono state selezionate solo quelle da cui, avendo subito una segmentazione, è stato quindi possibile estrarre i dati rilevanti a fini radiomici. Ciascuna osservazione è associata a 111 colonne, tra queste colonne, una rappresenta l'ID univoco per ciascun paziente, garantendo la tracciabilità e la corretta associazione dei dati. Inoltre, vi sono 3 colonne che rappresentano possibili variabili target, ovvero la variabile categorica che si vorrà predire.

La prima variabile candidata è la variabile qualitativa che descrive l'istologia del tumore, essa comprende tre categorie: Adenocarcinoma (ADN), Carcinoma a Cellule Squamose (SCC) e tipologia di tumore non specificata (NOS). L'adenocarcinoma è un tipo di tumore maligno che ha origine nelle cellule ghiandolari [15], mentre il Carcinoma a Cellule Squamose origina nelle cellule squamose, cellule epiteliali piatte e sottili [16].

La seconda variabile candidata è la variabile quantitativa binaria che essenzialmente raggruppa in due sottocampioni le osservazioni, il primo sottocampione comprende la diagnosi di ADN, il secondo quella di SCC o NOS.

Infine, la terza possibile variabile target è la recidività del tumore, un'altra variabile binaria che descrive dunque se il tumore è tornato dopo un periodo di remissione.

Nel nostro caso la variabile d'interesse era quella che forniva una diagnosi dell'istologia del tumore polmonare e in particolare la variabile binaria che avrebbe permesso di suddividere il dataset in due classi. Da un preliminare studio del dataset è infatti emerso che all'istologia di ADN erano associate 112 osservazioni, a quella di SCC 29, mentre a NOS, erano associate solo 3 osservazioni, queste due ultime istologie, come detto in precedenza, sono state unite all'interno della classificazione binaria.

Le restanti 107 colonne sono features, ovvero caratteristiche potenzialmente correlate con una delle variabili target, esse sono state estratte attraverso PyRadiomics [17], una piattaforma implementata nel linguaggio Python, flessibile e open-source capace di estrarre un ampio pannello di caratteristiche ingegnerizzate dalle immagini mediche. In particolare le features che possono essere estratte utilizzando PyRadiomics sono suddivise in classi in base alle seguenti caratteristiche[18]:

- Statistiche del Primo Ordine, esse descrivono la distribuzione delle intensità dei voxel all'interno della regione dell'immagine definita dalla maschera (regione d'interesse ROI), per esempio si ha l'energia legata ai voxel nell'immagine o la media dell'intensità del livello di grigio all'interno della ROI

- Forma 3D, ovvero descrittori delle dimensioni e della forma tridimensionale del ROI, basati quindi sulle caratteristiche dei voxel che caratterizzano questo, come per esempio sfericità o elongazione relativa del tumore
- Forma 2D, ovvero features analoghe alle precedenti, ma questa volta basate sulla forma bidimensionale e quindi sui pixel che compongono il ROI
- Matrice di co-occorrenza dei livelli di grigio, essa descrive la funzione congiunta di probabilità di secondo ordine del ROI, si ha per esempio la feature di autocorrelazione, che misura l'intensità della finezza della texture nella regione d'interesse o la feature che identifica il contrasto, ovvero la variazione locale dell'intensità di grigio
- Matrice di dimensione delle zone con lo stesso livello grigio
- Matrice di lunghezza delle serie con lo stesso livello di grigio
- Matrice di differenza di tonalità di grigio tra voxel vicini
- Matrice di dipendenza tra livelli di grigio vicini

1.2. Analisi Esplorativa

Il dataset in questione era già disponibile in un formato standardizzato e, infatti, il primo passaggio è stato quello di effettuare un'analisi esplorativa, essenziale per comprendere meglio la struttura e le peculiarità dei dati disponibili. Questo step è cruciale per identificare eventuali anomalie, pattern o trend nel dataset, permettendoci di pianificare in modo ottimale le successive fasi del nostro studio. Esso include un'analisi statistica e una visualizzazione dei dati e ha, inoltre, avuto il duplice scopo di ottenere informazioni utili sia per strutturare l'effettiva generazione di SD, che per confrontare, in secondo luogo, la qualità e la somiglianza dei dati creati rispetto a quelli reali. Per questo motivo, molti dei passaggi descritti nell'analisi esplorativa sono stati ripetuti sui dataset generati sinteticamente.

Quest'analisi è stata articolata in quattro step fondamentali:

- Data ingestion
- Studio della distribuzione della variabile target
- Studio delle features
- Analisi della correlazione tra features e variabile target

L'analisi è stata svolta, in particolare, servendosi del linguaggio Python nella versione 3.11.7 che offre diverse librerie e strumenti utili per generare SD nel contesto della radiomica.

1.2.1 Data Ingestion

Il primo passo dell'analisi esplorativa del dataset reale ha previsto un processo di strutturazione e archiviazione dei dati, noto come data ingestion. In questa fase, è stata dapprima identificata la variabile target di interesse, ovvero la variabile categorica che desideriamo predire. Nel nostro caso, la variabile target fornisce informazioni sull'istologia del carcinoma.

Successivamente, l'insieme delle osservazioni è stato suddiviso in due classi distinte in base alla variabile target:

- Nella prima classe sono state inserite le diagnosi di ADN
- Nella seconda classe, invece, sono state incluse le diagnosi di SCC e le

diagnosi di altre NOS.

Questa suddivisione permette di distinguere chiaramente tra le due categorie principali di interesse, agevolando così le successive fasi di analisi di dati.

1.2.2 Studio della distribuzione della variabile target

Il secondo passo dell'analisi esplorativa ha riguardato lo studio delle due classi precedentemente suddivise in base alla variabile target. In questa fase si è proceduti al calcolo e alla visualizzazione sia della frequenza assoluta che della frequenza relativa delle osservazioni all'interno di ciascuna classe.

Per ottenere una rappresentazione chiara e comprensibile dei dati, sono stati utilizzati grafici che hanno illustrato in modo efficace la distribuzione delle osservazioni. Infatti, questi hanno messo in risalto sia le differenze che le similitudini tra le due classi, permettendo così di sintetizzare in maniera efficace le informazioni, che altrimenti sarebbero state molto più complesse da individuare e comprendere.

1.2.3 Studio delle features

Il terzo passo ha previsto l'analisi delle distribuzioni delle features presenti nel dataset. Dopo aver verificato che non ci fossero dei valori mancanti in ogni osservazione, è stata effettuata un'analisi univariata, un'analisi bivariata e dei test d'ipotesi.

L'analisi univariata ha comportato lo studio delle singole variabili, al fine di comprenderne gli indici statistici fondamentali. L'analisi bivariata, invece, si è focalizzata sulle relazioni tra coppie di variabili, consentendo di individuare eventuali correlazioni e interazioni significative tra di esse.

Questi passaggi sono stati eseguiti inizialmente sull'intero dataset per ottenere una visione globale delle caratteristiche dei dati e, successivamente, le stesse analisi sono state applicate alle due classi precedentemente individuate e suddivise.

Gli indici statistici estratti sono:

- Media: la somma di tutti i valori di un insieme di dati divisa per il numero di

valori
- Mediana: il valore atteso della distribuzione
- Varianza: misura la dispersione dei valori rispetto alla media, più piccolo è il suo valore meno variabile sarà la distribuzione
- Curtosi: fa riferimento alla maggiore o minore gibbosità di una curva in prossimità del suo massimo e, quindi, alla maggiore o minore lunghezza delle code
- Asimmetria: è un termine che indica l'assenza di specularità di una distribuzione rispetto al suo asse di simmetria, per cui i valori del carattere di una distribuzione asimmetrica sono distribuiti con frequenze differenti attorno al suo valore centrale

In particolare, dato un dataset di n osservazioni e siano x_i i valori assunti da una certa variabile, gli indici statistici legati alla distribuzione di questa si calcolano come da Tabella 1.

Indice	**Formula**
Media	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Mediana	$\tilde{x} = x_{\left(\frac{n+1}{2}\right)}$ se n è dispari; $\tilde{x} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$ se n è pari
Varianza	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Curtosi	$\kappa = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$
Asimmetria	$\gamma = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$

Tabella 1. Formule degli indici statistici

Il campione è stato poi normalizzato; è stata, nello specifico, applicata la normalizzazione "min-max", ovvero una tecnica che ridimensiona i valori affinché questi rientrino in intervallo che va da 0 a 1 . Siano x_{min} e x_{max} rispettivamente il

valore minimo e massimo assunti da una distribuzione di valori x_i , allora i nuovi valori della distribuzione x'_i saranno dati dalla Formula (1.1).

$$x'_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{1.1}$$

Questo procedimento risulta molto vantaggioso nel momento in cui si applicano al dataset algoritmi basati sulla distanza tra i dati, poiché permette in particolare di ridurre gli errori di arrotondamento negli algoritmi numerici e in generale una convergenza più veloce degli algoritmi che utilizzano ML. Pertanto, gli algoritmi utilizzati per la generazione di SD saranno applicati al campione normalizzato, garantendo così prestazioni ottimali e una migliore gestione dei dati durante il processo di modellazione. Sul campione normalizzato sono stati successivamente calcolati gli stessi indici statistici precedentemente descritti e applicati al campione originale.

Tramite dei box plot e istogrammi è stato inoltre possibile visualizzare la distribuzione di ogni singola feature.

In seguito, è stata condotta un'analisi bivariata che ha confrontato la distribuzione delle singole features all'interno delle due stratificazioni del campione. Per visualizzare queste distribuzioni in modo chiaro, sono stati utilizzati boxplot e istogrammi. I boxplot hanno permesso di osservare la mediana, i quartili e la presenza di eventuali outlier per ogni feature. Gli istogrammi, d'altro canto, hanno mostrato la frequenza delle diverse categorie o intervalli di valori, facilitando il confronto visivo tra le due classi.

Parallelamente, per ottenere informazioni dettagliate sulle relazioni tra coppie di features, sono stati impiegati grafici di dispersione (scatter plot). Questi grafici hanno consentito di visualizzare eventuali correlazioni o pattern tra due variabili alla volta, mettendo in evidenza la distribuzione congiunta delle features nei gruppi di dati.

Successivamente, sono stati eseguiti vari test di normalità sulle features del dataset per valutare se le distribuzioni seguono una curva normale. Tra i test utilizzati, è stato incluso il test di Anderson-Darling, che si concentra specificamente sull'analisi delle code della distribuzione. Questo test è utile per rilevare deviazioni dalla normalità, soprattutto nelle estremità della distribuzione. Inoltre, è stato impiegato il

test di Kolmogorov-Smirnov, il quale verifica l'adattamento di un campione di dati a una distribuzione specifica, fornendo un'indicazione della conformità dei dati alla distribuzione teorica attesa. Un altro test utilizzato è stato quello di Shapiro-Wilk, noto per la sua efficacia nel rilevare deviazioni dalla normalità nei dati campionari.

Oltre ai test di normalità, è stato eseguito un independent T-test. Questo test statistico confronta le medie di due gruppi indipendenti per determinare se esiste una differenza significativa tra di esse. Nel nostro contesto, l'indipendent T-test è stato applicato ai due sottocampioni suddivisi in base all'istologia, per verificare se le differenze nelle medie delle features tra i due gruppi siano statisticamente significative.

Per tutti i test di ipotesi condotti, il livello di significatività, o p-value, è stato fissato a 0.05. Questo significa che le differenze osservate sono considerate statisticamente significative se il p-value ottenuto è inferiore a 0.05.

1.2.4 Analisi della correlazione tra features e variabile target

Il quarto e ultimo passo dell'indagine esplorativa consiste nell'analizzare la correlazione tra tutte le diverse coppie di features presenti nel dataset, nonché tra ciascuna feature e la variabile target. Per questa analisi, è stato utilizzato il coefficiente di correlazione di Pearson, uno strumento statistico che consente di quantificare la correlazione lineare tra due variabili. Il coefficiente di Pearson fornisce un valore che varia tra -1 e 1, dove un valore di 1 indica una correlazione positiva perfetta, un valore di -1 indica una correlazione negativa perfetta, e un valore di 0 indica che non esiste alcuna correlazione tra le variabili. Per visualizzare meglio i risultati abbiamo ritenuto utile rappresentare una heat map che mostra la matrice di correlazione che compara le coppie di features.

Sono state poi selezionate le features altamente correlate impostando un valore di soglia pari a 0.95. Un’alta correlazione può essere infatti sintomo di collinearità, essa dunque determina la presenza di informazioni ridondanti che comprometterebbe l’algoritmo da allenare con il dataset in questione; è comunque da tener presente che un’alta correlazione tra due variabili non è sempre legata ad un rapporto causa-effetto. Potrebbe infatti sussistere il fenomeno del “co-founding”, dove la

relazione osservata tra due variabili è dovuta all'influenza di una terza variabile non considerata.

Dopo aver analizzato la correlazione tra le features e la variabile target, sono state selezionate quelle features con una bassa correlazione, impostando una soglia a 0.01. Questo processo è stato adottato per identificare e rimuovere le variabili che non apportano un valore aggiunto significativo al modello predittivo. Per generare i SD sarà necessario garantire un somiglianza delle correlazioni.

1.3. Pre-processing

Il pre-processing è un processo composto da una serie di passaggi che assicura che i dati siano correttamente formattati per essere forniti in input agli algoritmi che saranno utilizzati in seguito, migliorando così la loro efficacia e accuratezza.

I passaggi eseguiti durante il pre-processing, a partire dal dataset originale, sono stati:

- Divisione del dataset in base alla variabile target, creando due sottocampioni distinti
- Normalizzazione dei sottocampioni, con ridimensionamento dei valori delle features su una scala comune migliorando la performance degli algoritmi basati sulla distanza
- Eliminazione delle features ad alta correlazione, cioè in presenza di collinearità, una delle due features altamente correlate è stata eliminata, al fine di evitare la ridondanza informativa
- Eliminazione delle features a bassa correlazione con la variabile target, che non appartenevano un contributo significativo alla predizione.

Questo doppio processo di selezione, basato sull'identificazione sia delle alte correlazioni tra le features che delle basse correlazioni con la variabile target, ha garantito che le variabili rimanenti nel dataset fossero sia rilevanti che non ridondanti.

1.4. Metodi di Generazione

1.4.1 Algoritmo Vale & Maurelli

Per l'effettiva generazione del dataset sintetico di features radiomiche si è optato per utilizzare due metodi e confrontare i risultati ottenuti. Il primo metodo utilizzato si può classificare all'interno dei metodi tradizionali ed è l'algoritmo implementato da Vale & Maurelli (VM) [12], esso è basato sul metodo Monte Carlo (MC) e viene usato per generare distribuzioni multivariate non normali simulando un processo casuale o aleatorio. Le principali caratteristiche dei metodi MC sono tre, ovvero: generazione di numeri casuali, esecuzione di una simulazione in questione per molte iterazioni e l'aggregazione dei risultati delle singole simulazioni per ottenere una media statistica.

Il metodo VM, come detto in precedenza, viene utilizzato quando si desidera generare campioni di SD che non seguono una distribuzione normale ma che mantengono certe proprietà statistiche. In particolare, all'interno dell'algoritmo, si calcola la variabile non normale Y come combinazione lineare delle prime tre potenze di una variabile casuale normale standard X, come da Formula (1.2).

$$Y = a + bX + cX^2 + d\,X^3 \qquad (1.2)$$

Le costanti a, b, c e d sono i coefficienti di Fleishman, esse sono state scelte per fornire a Y una distribuzione specifica a partire dagli indici statistici calcolati su dati reali forniti in input.

Nello specifico l'algoritmo per la generazione di SD tabulari è stato implementato attraverso i seguenti passaggi:

1. Calcolo degli indici statistici di media, varianza, asimmetria e curtosi per ciascuna variabile nel dataset originale
2. Calcolo dei coefficienti di Fleishman a partire dagli indici statistici appena calcolati
3. Calcolo della matrice di correlazione intermedia a partire dalla matrice di correlazione tra le diverse variabili del dataset fornito e i coefficienti di Fleishman appena calcolati. Si verifica poi che questa matrice sia definita positiva

4. Generazione di un campione X di variabili normali multivariate a partire dalla matrice di intercorrelazione precedentemente determinata
5. Applicazione della trasformazione non lineare in Formula (1.2) per il calcolo del campione Y di variabili non normali.

L'algoritmo VM per la generazione di campioni multivariati non normali è semplice e veloce ma possiede dei limiti teorici, in particolare per certe combinazioni di curtosi e asimmetria, è dunque necessario confrontare le proprietà statistiche dei SD con quelle dei dati reali per assicurarsi che effettivamente corrispondano. Quindi la qualità dei dati reali utilizzati per identificare le distribuzioni empiriche può influenzare direttamente la qualità dei SD generati. Dati reali non rappresentativi o con errori significativi possono compromettere l'accuratezza dei SD. D'altro canto se i dati reali mostrano una variabilità complessa in certi casi non può essere completamente catturata.

1.4.2. Metodo CTGAN

Il secondo metodo, invece, è stato scelto all'interno della categoria dei metodi di DL, in particolare dalla ricerca bibliografica è emerso che il metodo più adatto alla nostra caso fosse il CTGAN: il metodo GAN è una tecnica di apprendimento automatico nota come rete neurale generativa avversaria e nello specifico il metodo CTGAN viene utilizzato prevalentemente per dati eterogenei tabulari (come nel nostro caso). [13],[19]

Questo algoritmo è formato da due reti che svolgono due ruoli diversi:

- Il generatore (*Generator*): Questa parte del modello è responsabile della creazione di SD. Prende come input un vettore di rumore casuale (spesso chiamato "rumore latente") e lo trasforma in un'immagine (o altro tipo di dato) che assomiglia a quelle nel set di dati di addestramento.
- Il discriminatore (*Discriminator*): Questa parte del modello è responsabile della distinzione tra dati reali e SD. Prende in input un'immagine (o altro tipo di dato) e produce una previsione su quanto sia probabile che l'input sia reale piuttosto che sintetico.

Lo scopo delle due reti è quello di competere tra di loro, in modo da migliorarsi continuamente.

Il processo di addestramento di una rete GAN, di cui ogni ciclo è definito epoch, è composto da due fasi in competizione tra loro:

- Fase di addestramento del discriminatore: In questa fase, il discriminatore viene addestrato utilizzando un set di dati contenente sia esempi reali che sintetici e il suo compito è distinguere accuratamente tra i due. Il generatore è disattivato durante questa fase.
- Fase di addestramento del generatore: In questa fase, il generatore viene addestrato per generare dati che ingannino il discriminatore, cercando di produrre SD che siano così realistici da essere classificati dal discriminatore come dati reali. Il discriminatore viene quindi utilizzato per valutare quanto validi siano i dati prodotti dal generatore.

Queste due fasi si alternano in modo iterativo durante il processo di addestramento finché il generatore non riesce a generare SD che sono indistinguibili dai dati reali secondo il discriminatore. Il numero di iterazioni determina quanto a lungo il modello verrà addestrato. In generale, un numero maggiore di epoch consente al modello di apprendere meglio, ma può aumentare il rischio di overfitting, dove il modello si adatta troppo ai dati di addestramento e al loro rumore e non generalizza bene i nuovi dati; al contrario poche epoch possono causare underfitting, dove il modello non apprende abbastanza dai dati di addestramento generando dati poco variabili e non rappresentativi delle relazioni esistenti.

Abbiamo scelto il metodo CTGAN, poiché dalla ricerca bibliografica è risultato essere in grado di mantenere la complessità e le dipendenze dei dati, gestire lo squilibrio delle classi, e produrre SD di alta qualità. Il metodo è inoltre relativamente recente e sta rapidamente guadagnando sempre di più attenzione da parte della comunità scientifica, questo implica un crescente numero di miglioramenti, adattamenti e documentazioni che possono facilitare l'implementazione e l'ottimizzazione del metodo per esigenze specifiche.

1.5. Quality evaluation e calcolo dell'errore relativo

Infine, per ogni metodo si applica una quality evaluation, un'analisi che permette di verificare che i dati generati siano validi (i.e. id univoci e autentici, SD rispettano i valori max e min imposti dai dati reali, ecc.). La quality evaluation è stata effettuata utilizzando la libreria SDV(Synthetic Data Vault), una libreria implementata in Python pensata per generare dati tabulari sintetici per verificare la correttezza dei dati generati.

In primo luogo si fa il controllo delle metriche data validity e data structure. La data validity permette di assicurarsi che i SD siano affidabili, e quindi capaci di rappresentare correttamente le caratteristiche dei dati reali, riducendo al minimo i bias. Essa è valutata attraverso la boundary adherence, che si riferisce alla capacità dei SD di rispettare i limiti e i vincoli presenti nei dati originali. Nella data structure invece si verifica che la rappresentazione dei SD rispecchia correttamente la struttura dei dati reali. Essa è valutata attraverso la table structure, che verifica che la disposizione dei SD sia in formato tabellare, analogamente ai dati reali.

Si procede poi con due altre valutazioni: le column shapes e i column pair trends. Le column shapes verificano se i dati fittizi hanno le stesse proprietà di quelli originali e vengono misurate calcolando il KSC complement, che è una metrica utilizzata per valutare la similarità tra due distribuzioni di probabilità. I column pair trends invece valutano le relazioni tra coppie di colonne nei SD rispetto a quelli reali, si misurano calcolando la correlation similarity, una metrica utilizzata per valutare quanto due insiemi di dati siano correlati tra loro.[20]

Oltre alla quality evaluation, per poter misurare e mettere a confronto i due gruppi di SD generati per ogni metodo, è stato calcolato l'errore relativo degli indici statistici di ogni feature (media, varianza, asimmetria e curtosi). Successivamente è stata utilizzata la norma di Frobenius per confrontare le matrici di correlazione vera e generata sinteticamente per ogni metodo.

L'errore relativo è stato calcolato come l'errore relativo medio totale ottenuto aggregando tutti gli errori relativi per ciascun indicatore statistico per ciascuna feature e confrontando i dati reali e i dati generati sinteticamente. In particolare, ogni indice statistico rappresenta una misura particolare, come la media, la varianza o un altro momento statistico, che descrive una particolare feature dei dati. Questo processo è stato eseguito proprio per valutare l'accuratezza dei SD generati rispetto ai

dati originali. Per poter implementare questa valutazione, i passaggi sono stati eseguiti separatamente per i due metodi di generazione di SD scelti: il metodo VM e il CTGAN. Mentre il metodo di VM è stato testato una volta, il metodo CTGAN è stato valutato in due diverse configurazioni, corrispondenti rispettivamente a 300 e 600 epochs. Utilizzando diversi insiemi di epochs, siamo stati in grado di analizzare come il numero di iterazioni influisce sulla qualità dei SD generati.

L'obiettivo principale di questo approccio è quantificare l'errore tra la distribuzione marginale dei SD e la distribuzione marginale dei dati reali. Le distribuzioni marginali rappresentano le probabilità che diverse features appaiano in modo indipendente e, confrontandole, è possibile determinare quanto bene i SD riflettono le features dei dati originali. Questo processo dettagliato di calcolo e confronto degli errori relativi fornisce una misura quantitativa della somiglianza tra SD e reali e consente determinare quali metodi di generazione di SD possono produrre risultati che abbiano una migliore similarità delle features del nostro dataset vero.

Per rafforzare ulteriormente questa valutazione, abbiamo utilizzato la norma di Frobenius come metrica per misurare la similarità tra le matrici di correlazione sintetiche e quelle reali. La norma di Frobenius è una misura della "dimensione" di una matrice ed è una delle norme più comunemente utilizzate nell'algebra lineare. Formalmente, la norma di Frobenius di una matrice A di dimensione m x n è definita come la radice quadrata della somma dei quadrati di tutti gli elementi della matrice .

Nel nostro caso, per ogni metodo che abbiamo utilizzato, si calcola la differenza tra la matrice di correlazione reale e la matrice generata sinteticamente. In questo modo si crea una nuova matrice di differenza (Formula (3)). Ogni elemento della differenza delle matrici rappresenta la differenza tra gli elementi corrispondenti della matrice originale. La norma di Frobenius viene quindi calcolata come radice quadrata della somma dei quadrati di tutti questi elementi di differenza (Formula (4)). Un valore inferiore alla norma indica un'elevata somiglianza tra la matrice di correlazione effettiva e la matrice generata sinteticamente.

Definiamo la matrice D, dove come la differenza tra le matrici di correlazione, dove A_{vero} corrisponde alla matrice di correlazione dei dati veri, e A_{synt} corrisponde alla matrice di correlazione sinteticamente generata:

$$D = A_{\text{vero}} - A_{\text{synt}} \tag{3}$$

Utilizziamo la norma di Frobenius con la matrice D, ottenuta nella Formula (3):

$$\|D\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|d_{ij}|^2} \tag{4}$$

Può essere applicato per valutare quanto si avvicina la matrice di correlazione dei SD alla matrice di correlazione dei dati reali. Questo confronto è molto importante perché la matrice di correlazione rappresenta le relazioni che si presentano tra le diverse features dei dataset e il mantenimento di queste relazioni è essenziale per garantire che i SD possano rappresentare in modo adeguato i dati originali.

2 Risultati

2.1. Risultati dell'analisi esplorativa

I risultati ottenuti dall'analisi esplorativa sono stati fondamentali per impostare correttamente il processo di generazione di SD. Essi sono stati organizzati in modo tale da mostrare i risultati di ogni passaggio dell'analisi esplorativa. Quindi, abbiamo diviso i risultati in:

- Risultati del data ingestion
- Risultati dell'analisi univariata e analisi bivariata
- Risultati dei test statistici
- Risultati dell'analisi di correlazione

2.1.1 Risultati del data ingestion

Nella fase di data ingestion abbiamo diviso l'intera popolazione in 2 classi a seconda della presenza istologica tumorale di ogni osservazione, una popolazione con la presenza di adenocarcinoma (ADN) e un'altra con la presenza di cellule squamose o altri tumori non specificati (SCC o NOS). In particolare è stato rilevato uno sbilanciamento tra le due popolazioni, come viene visualizzato in Figura 1. Infatti, abbiamo evidenziato come il calcolo delle frequenze relative delle due popolazioni comprendeva le osservazioni a cui corrispondeva la diagnosi di ADN, era pari al 77.8%, quindi significativamente più alta del sottocampione che comprendeva diagnosi di SCC o NOS con una percentuale del 22.2%.

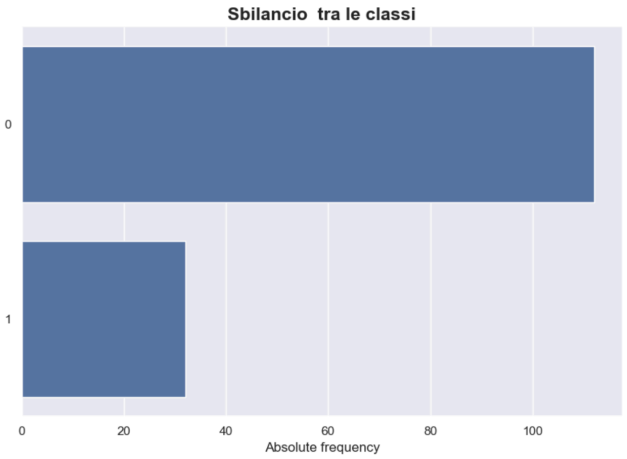

Figura 1 : Frequenza assoluta tra le due popolazioni. Popolazione ADN = 0 Popolazione SCC/NOS = 1

Questo sbilanciamento tra le classi può portare a dei problemi nell'addestramento di algoritmi, che infatti tenderanno a favorire la classe con più occorrenze. Lo sbilanciamento è stato risolto costruendo 2 generatori di SD per ogni classe in modo da avere due popolazioni bilanciate. Inoltre, abbiamo verificato che non ci fossero dei valori mancanti in nessuna delle due popolazioni.

2.1.2 Risultati dell'analisi univariata e bivariata

Nell'analisi univariata, la prima cosa che abbiamo fatto è stata quella di raccogliere i valori della media, la varianza, la skewness e la curtosi della distribuzione marginale di ogni feature. A titolo esemplificativo, abbiamo riportato i grafici di una sola feature, un boxplot che ci ha permesso di identificare i valori interquartili di ogni feature e un istogramma per visualizzare la distribuzione marginale di ogni feature. I valori sono stati normalizzati per facilitare il confronto tra le feature e diminuire la propagazione degli errori di arrotondamento. L'analisi univariata è stata fatta per l'intera popolazione.

Abbiamo riportato in Figura 2a il boxplot che visualizza la distribuzione sull'intero campione della feature "*original_shape_Elongation*", si nota chiaramente che la distribuzione non è simmetrica, poiché presenta una coda sinistra più lunga. Inoltre, notiamo che c'è anche un outlier a sinistra. La gestione degli outliers è importante e

devono essere trattati con attenzione, perché in alcuni casi portano informazioni utili e veritiere, mentre in altri casi sono solo errori di misurazione. Gli outliers possono essere identificati con la misura z-score per poi essere rimossi. Invece, per quanto riguarda la generazione di SD sono stati considerati anche gli outliers perché l'obiettivo dei SD era quello di rispettare la struttura interna e le proprietà statistiche interne dei dati originali considerando anche i casi rari. Le informazioni derivanti dall'analisi possono essere confermate attraverso la rappresentazione di un istogramma, come mostrato nella Figura 2b.

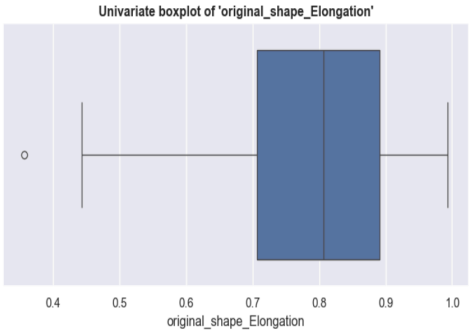

Figura 2a:

Boxplot della feature 'original_shape_elongation'

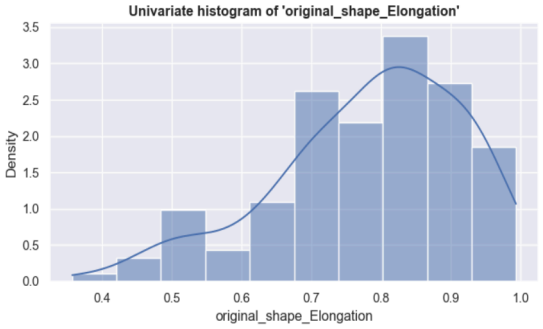

Figura 2b:

Istogramma della feature 'original_shape_elongation'

Nell'analisi bivariata, analogamente alla procedura fatta nell'analisi univariata, dove abbiamo individuato le misure statistiche di ogni feature, in quest'analisi abbiamo confrontato la distribuzione marginale di ogni feature per ogni popolazione visualizzandole tramite boxplots e istogrammi. A titolo d'esempio abbiamo riportato due features, la feature menzionata prima, *"original_shape_Elongation"* e la feature *"original_glcm_Contrast"* .

Confrontando la distribuzione delle singole features nei due sottocampioni, è emerso che alcune distribuzioni erano notevolmente simili tra la popolazione con presenza di ADN e la popolazione con presenza di SCC/NOS, come chiaramente illustrato dal boxplot (Figura 3a) e dall'istogramma (Figura 3b) della feature "*original_shape_Elongation*".

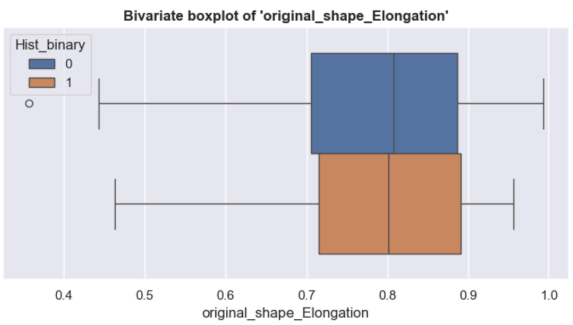

Figura 3a:

Boxplot della feature *"originale_shape_Elongation"*. Popolazione ADN(in blu) Popolazione SCC/NOS(in arancione)

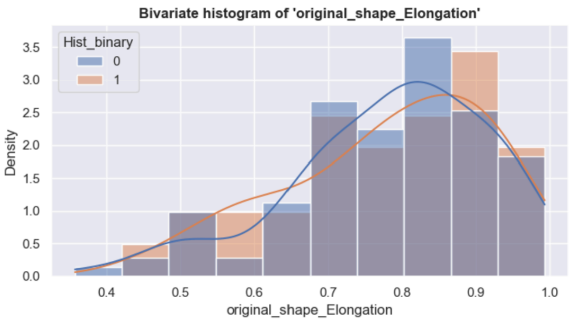

Figura 3b:

Istogramma della feature *"originale_shape_Elongation"*. Popolazione ADN(in blu) Popolazione SCC/NOS(in arancione)

Questa somiglianza indica una coerenza significativa nella distribuzione di questa feature tra i due gruppi di dati. In Figura 3a si nota come la mediana di entrambe le distribuzione sia circa 0.8 indicando una asimmetria nelle sue distribuzioni marginali, come viene visualizzato in Figura 3b.

Tuttavia, in altri casi le distribuzioni erano piuttosto differenti, come evidenziato nella Figura 4a, che mostra il boxplot della feature *"original_glcm_Contrast"*. Qui, le differenze marcate nella distribuzione sottolineano variazioni sostanziali tra i due sottocampioni. In Figura 4b, si confrontano distribuzioni marginali tra le due popolazioni tramite gli istogrammi.

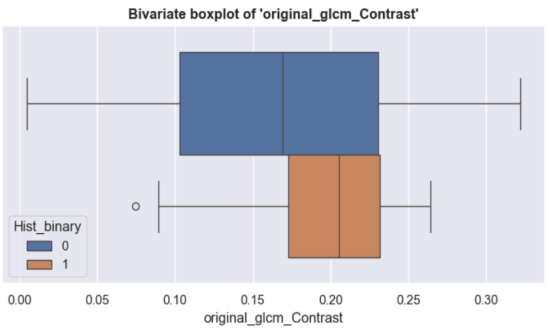

Figura 4a:

Boxplot della feature "*originale_glcm_Contrast*". Popolazione ADN(in blu) Popolazione SCC/NOS(in arancione)

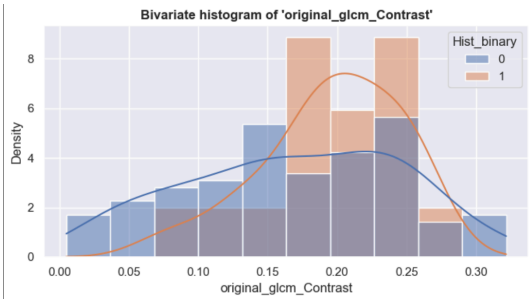

Figura 4b:

Istogramma della feature "*originale_glcm_Contrast*". Popolazione ADN(in blu) Popolazione SCC/NOS(in arancione)

2.1.3 Risultati test statistici

Attraverso i test di normalità eseguiti, poi, è emerso che analizzando le distribuzioni delle altre feature presenti nel dataset, la maggior parte di esse non segue una distribuzione normale. Questa caratteristica non solo definisce la natura non parametrica dei dati originali (non seguono infatti una determinata distribuzione e sarà necessario utilizzare test statistici non parametrici), ma rappresenta anche una considerazione fondamentale da integrare nella generazione di SD. È essenziale che i SD replichino accuratamente questa non normalità, mantenendo così la coerenza con le caratteristiche del dataset originale. Quindi, alla fine della generazione dei SD abbiamo fatto un test di normalità per i dati SD, dove abbiamo messo come ipotesi nulla che le distribuzioni marginali fossero normali.

Inoltre, l'analisi mediante l'independent T-test ha rivelato che, in generale, le medie delle due distribuzioni nelle popolazioni con ADN e con SCC/NOS erano molto simili per quasi tutte le features.

2.1.4 Risultati dell'analisi di correlazione

Infine, abbiamo proceduto con un'analisi correlativa tra features da cui è emersa una

matrice di correlazione tra le coppie di features con alta complessità e diversificazione, come evidenziato anche dalla heatmap mostrata in Figura 5 che visualizza questa matrice. A destra, viene mostrata una colorbar che indica i valori del coefficiente di Pearson tra le features da 0 ad 1 dove, un numero vicino ad uno indica un'alta correlazione tra due features, mentre valori vicini a zero indicano una bassa correlazione tra due features.

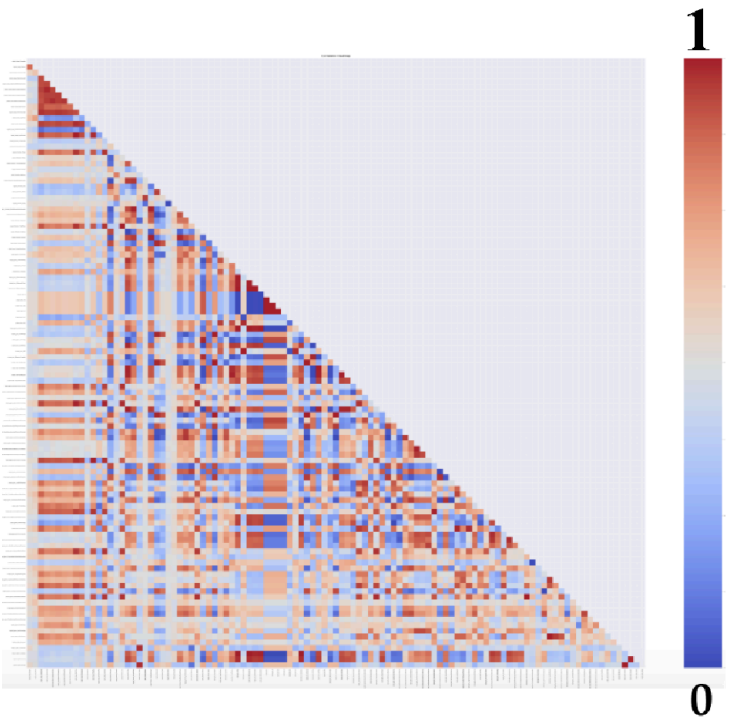

Figura 5 : Heatmap della matrice di correlazione delle features

È stato poi riscontrato che in tutta la popolazione erano presenti 121 coppie di variabili con un'alta correlazione mentre nel caso della popolazione con ADN erano 140 e nel caso della popolazione con SCC/NOS erano 141. Questo ci porterà in seguito a effettuare la rimozione delle features altamente correlate tra di loro per ogni classe di popolazione con lo scopo di evitare una eventuale collinearità tra variabili che causerebbe dei problemi nel momento in cui si applica l'algoritmo per la generazione dei SD.

Per quanto riguarda invece lo studio della correlazione delle features con la variabile target è stato riscontrato un valore massimo in modulo di 0.3 e in generale i valori dei coefficienti sono distribuiti quasi uniformemente nell'intervallo (-0.3, 0.3), come si

evince dalla Figura 6, si parla quindi di una correlazione debole. Questo suggerisce che non ci sia una feature con una dominanza assoluta rispetto alle altre.

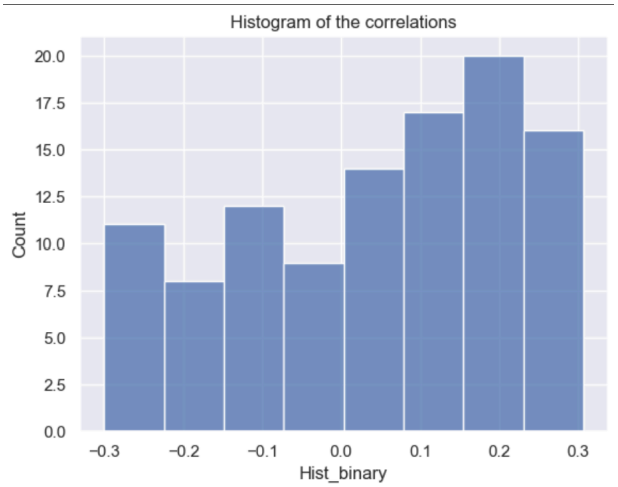

Figura 6 : Istogramma delle correlazioni tra le features e il target

2.2. Risultati del metodo Vale & Maurelli

La prima caratteristica che si può evidenziare nell'applicazione dell'algoritmo VM è che la procedura di Fleishman per la produzione di distribuzioni di variabili non permette di conoscerne l'esatta distribuzione e quindi manca di funzioni di densità di probabilità e di distribuzione cumulativa e dunque, inoltre, non può produrre distribuzioni con tutte le possibili combinazioni di asimmetria e curtosi come riscontrato anche in letteratura. [12]

Dopo aver effettuato la generazione del dataset sintetico con il metodo di VM sono stati calcolati e rappresentati graficamente gli indici statistici di base, di cui di seguito si riportano i grafici relativi alla media, che risultano essere i più rilevanti a livello informativo. Nel grafico rappresentato in Figura 7 è presente, appunto, la media delle features reali a confronto con la media di quelle generate sinteticamente, in questo caso era stato utilizzato l'interno dataset per la generazione. In Figura 8 e Figura 9 si ha lo stesso confronto descritto in precedenza ma con una generazione basata rispettivamente sul primo e secondo sottocampione.

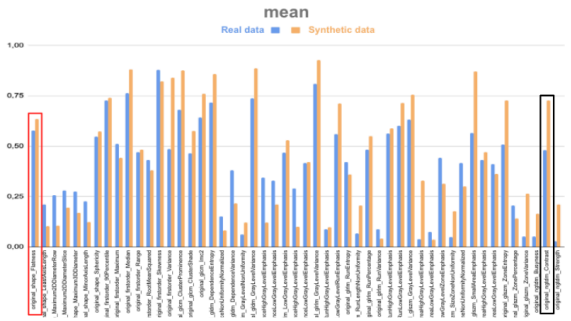

Figura 7 : Media delle features generate con V&M con il campione completo

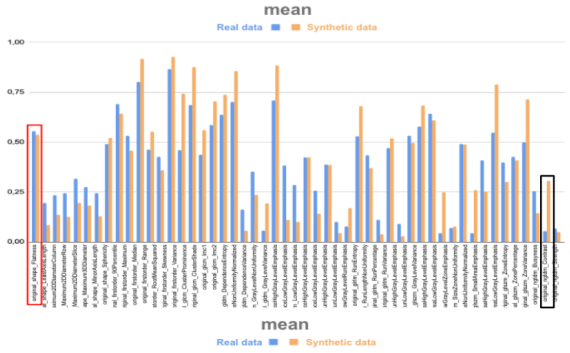

Figura 8 : Media delle features generate con V&M con il primo sottocampione (ADN)

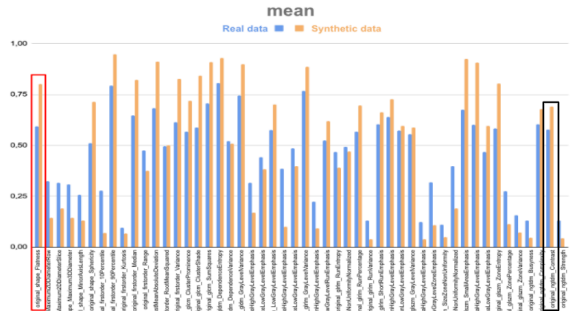

Figura 9 : Media delle features generate con V&M con il secondo sottocampione (SCC o NOS)

Per effettuare un'analisi più specifica è stato utile guardare nello specifico il comportamento di due features, le quali sono state evidenziate anche nelle immagini precedenti: la prima descrive quanto è piatta la forma del tumore, ovvero la feature "*original_shape_Flatness*" e la seconda descrive il contrasto tra le zone di grigio vicine, ovvero la feature "*original_ngdtm_Contrast*". Queste due features sono rispettivamente una delle meglio riprodotte e una delle peggio riprodotte dai SD e saranno confrontate anche per il dataset generato con il secondo metodo.
Per studiare meglio le distribuzioni delle diverse variabili sono stati utilizzati dei boxplot.

Per quanto riguarda la prima feature, ovvero quella che rappresenta la piattezza del tumore ("*original_shape_Flatness*") possiamo notare che:

- Nella popolazione totale (Figura 10) lo scarto interquartile (IQR) risulta simile a livello della mediana e del terzo quartile, ma il primo quartile è leggermente diverso. Gli outlier sono disposti diversamente nei SD, quindi abbiamo valori estremi diversi.

- Nella popolazione con ADN (Figura 11) si ha, invece, una situazione dei SD molto più realistica: infatti, abbiamo scarto interquartile molto simile e primo quartile, mediana e terzo quartile vicini a quelli reali. Inoltre, non abbiamo outliers né nei SD né in quelli reali. Tutte queste caratteristiche suggeriscono quindi una riproduzione dei dati affidabile.

- Nella popolazione con SCC/NOS (Figura 12) i SD non riescono più a riprodurre la realtà in modo molto attendibile. Infatti, pur essendo lo scarto interquartile (IQR) di dimensione simile, sono molto diversi il primo quartile, la mediana e il terzo quartile. L'estremo inferiore è molto spostato nei SD e gli outliers sono disposti in maniera molto diversa.

Tutte queste caratteristiche suggeriscono quindi una riproduzione dei dati non fedele.

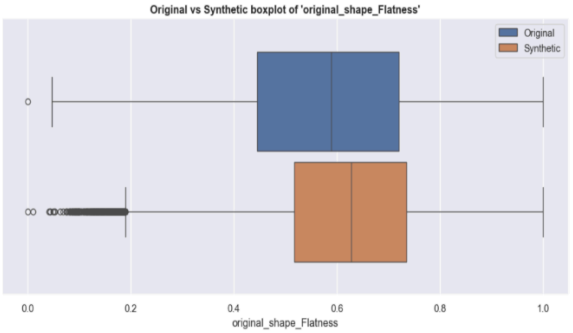

Figura 10 : Distribuzione feature "*original_shape_Flatness*" del campione complete

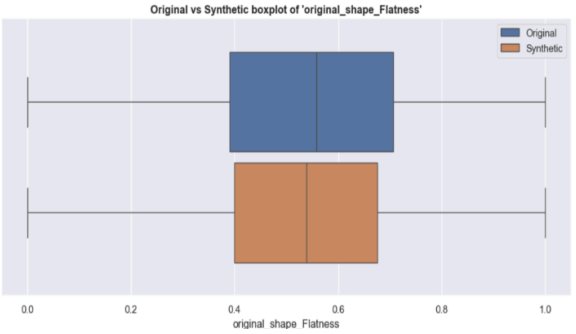

Figura 11 : Distribuzione feature "*original_shape_Flatness*" del primo sottocampione (ADN)

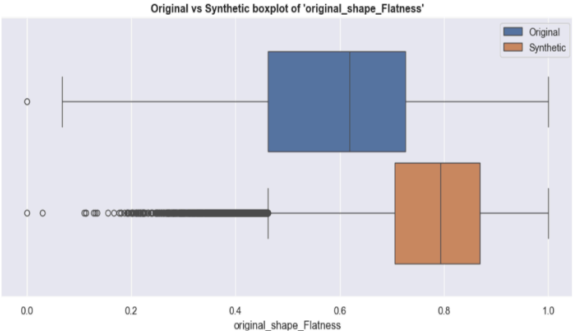

Figura 12 : Distribuzione feature "*original_shape_Flatness*" del secondo sottocampione (SCC o NOS)

Per quanto riguarda la seconda feature, ovvero quella che rappresenta il contrasto ("*original_ngdtm_Contrast*") possiamo invece notare che:

- Nella popolazione totale (Figura 13) abbiamo scarto interquartile discostato tra dati fittizi e reali di circa 0.2 e così anche il primo quartile, la mediana e il terzo quartile. L'estremo inferiore è molto maggiore nel caso dei SD, ma l'estremo superiore coincide. Per quanto riguarda gli outliers, invece, nei SD ne abbiamo uno in corrispondenza dell'estremo inferiore dei dati reali.

- Nella popolazione con ADN (Figura 14) notiamo invece un miglioramento nella mediana, che questa volta coincide, ma rimangono molto diversi lo scarto interquartile, il primo e il terzo quartile. In questo caso l'estremo inferiore e superiore dei SD coincidono quasi con il primo e il terzo quartile dei dati reali. Inoltre, ci sono molti outliers nei SD, mentre nei dati reali sono assenti.

- Nella popolazione con SCC/NOS (Figura 15) si può notare un leggero miglioramento in quanto la dimensione dello scarto interquartile risulta più simile, tuttavia primo quartile, mediana e terzo quartile non coincidono. Si hanno però l'estremo superiore e inferiore molto simili a quelli dei dati reali e anche gli outliers si trovano circa nella stessa zona, sebbene non coincidano in termini quantitativi.

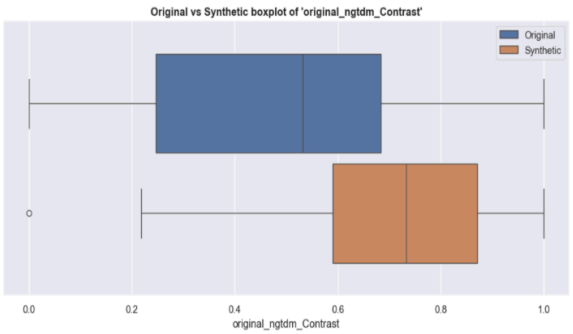

Figura 13 : Distribuzione feature *"original_ngdtm_Contrast"* del campione completo

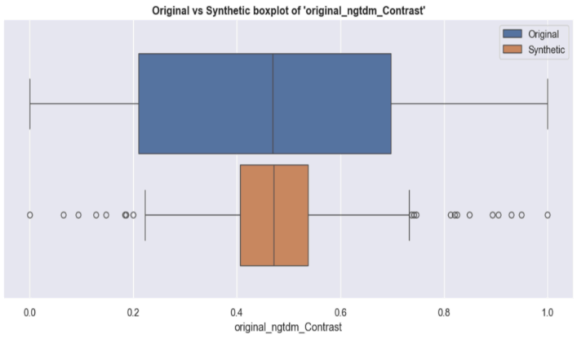

Figura 14 : Distribuzione feature *"original_ngdtm_Contrast"* del primo sottocampione (ADN)

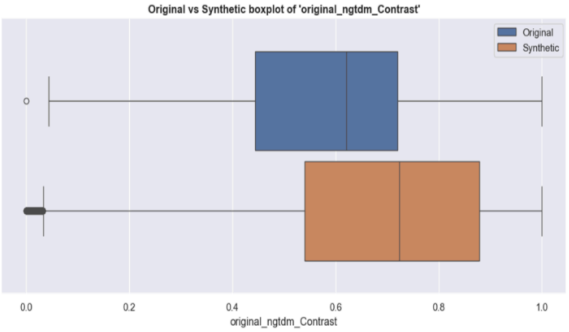

Figura 15 : Distribuzione feature "*original_ngdtm_Contrast*" del secondo sottocampione (SCC o NOS)

Una delle caratteristiche fondamentali per cui il dataset sintetico doveva rispecchiare quello reale era la correlazione tra le diverse feature.

Feature 1	**Feature 2**	**Correlazione RL**	**Correlazione SL**
original_shape_Elongation	original_shape_MajorAxisLength	-0.222	-0.189
original_ngtdm_Coarseness	original_ngtdm_Strength	0.999	0.091

Tabella 2 Coefficiente di correlazione di Pearson tra due coppie di features nel dataset reale (RL) rispetto a quello sintetico (SL)

Dall'analisi della correlazione tra le diverse features è risultato che, se per molte di queste il coefficiente di correlazione di Pearson è molto simile, come riporta l'esempio nella prima riga della Tabella 2, per altre coppie di variabili la correlazione era completamente differente nei dati appena generati con il metodo di VM rispetto a quella nel dataset di partenza. Come si vede nella secondo riga della Tabella 2, infatti, se la feature "*original_ngtdm_Coarseness*" e "*original_ngtdm_Strength*" avevano praticamente una proporzionalità diretta, la coppia di variabili nei dati generati risulta avere una correlazione molto bassa: è evidente come in questa caratteristica la feature generate non rispecchiano quelle reali. Questo potrebbe essere legato al fatto

che il metodo utilizza modelli matematici/statistici semplici per generare i SD, che potrebbero non catturare tutte le complessità e le relazioni presenti nei dati reali. Ad esempio, potrebbe trattare le relazioni tra le variabili in modo lineare mentre le relazioni reali possono non esserlo.

2.3. Risultati del metodo CTGAN

Anche per il metodo del CTGAN per ogni feature sono stati calcolati i vari indici statistici, e in particolare è stato realizzato un grafico della media per ogni gruppo (popolazione totale: Figura 16; ADN: Figura 17; SCC/NOS: Figura 18) mettendo a confronto quella del dato originale e del dato generato sinteticamente. Sono state identificate due istanze, evidenziandole sui grafici: la piattezza del tumore, ovvero la feature "*original_shape_Flatness*" che è una tra le meglio rappresentate, e il contrasto tra le zone di grigio vicine, ovvero la feature "*original_ngdtm_Contrast*" che è invece una tra le peggio rappresentate. Per ciascuna è stato realizzato un boxplot da cui si possono ricavare più informazioni.

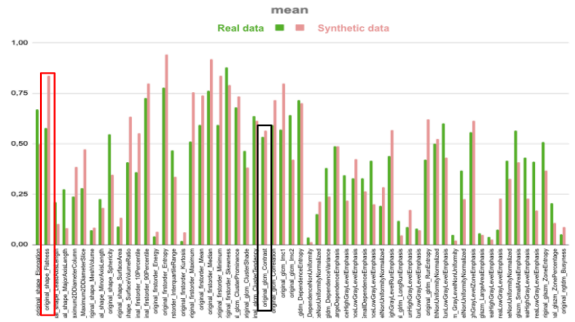

Figura 16 : Media delle features generate con CTGAN con il campione completo

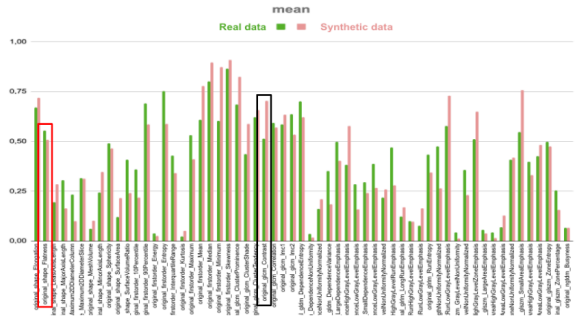

Figura 17 : Media delle features generate con CTGAN con il primo sottocampione (ADN)

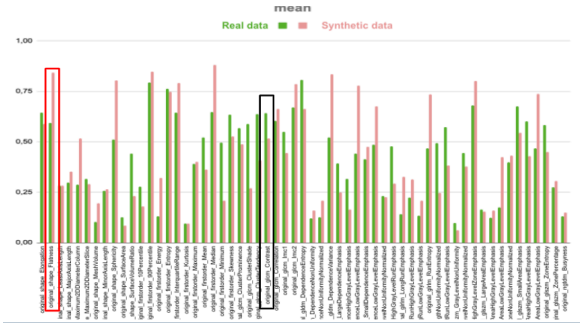

Figura 18 : Media delle features generate con CTGAN con il secondo sottocampione (SCC o NOS)

Per la prima feature (piattezza del tumore) si può notare:

- Nella popolazione totale (Figura 19) si ha lo scarto interquartile (IQR) molto simile; il terzo quartile dei SD è vicino a quelli dei dati reali, ma il primo quartile e la mediana sono leggermente diversi.. Gli outlier presenti nei dati originali non sono riprodotti nei SD quindi si hanno estremi diverse.

- Nella popolazione con ADN (Figura 20) la situazione dei SD risulta molto più fedele: infatti, si ha scarto interquartile praticamente identico, primo e terzo quartile vicini a quelli reali. La mediana risulta leggermente discostata tra i dati reali e sintetici. Infine , non abbiamo outliers né nei SD né in quelli reali.

- Nella popolazione con SCC/NOS (Figura 21), pur essendo lo scarto interquartile (IQR) di dimensione simile, sono molto diversi il primo quartile, la mediana e il terzo quartile. Oltretutto, gli outliers presenti nei dati originali non sono riprodotti in quelli sintetici.

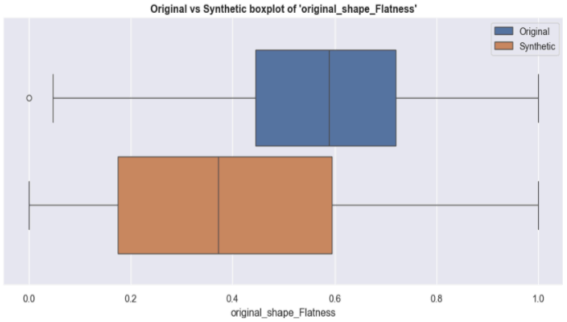

Figura 19 : Distribuzione feature "*original_shape_Flatness*" del campione completo

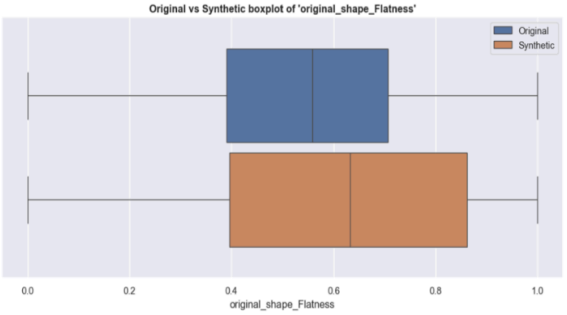

Figura 20 : Distribuzione feature "*original_shape_Flatness*" del primo sottocampione (ADN)

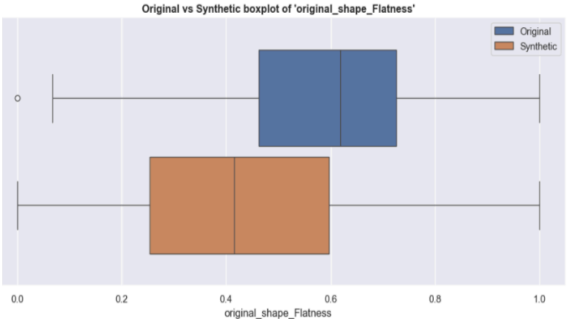

Figura 21 : Distribuzione feature "*original_shape_Flatness*" del secondo sottocampione (SCC o NOS)

Per quanto riguarda la seconda feature (contrasto) possiamo invece notare:

- Nella popolazione totale (Figura 22) si ha scarto interquartile e mediana simili. Non sono presenti in entrambi i casi outliers.
- Nella popolazione con ADN (Figura 23) lo scarto interquartile continua a coincidere, ma con un evidente peggioramento degli estremi, infatti si nota che la mediana dei SD coincide quasi con l'estremo inferiore dei dati originali; mentre la mediana dei dati originali è quasi allo stesso livello dell'estremo superiore dei SD.
- Nella popolazione con SCC/NOS (Figura 24) si registra un ulteriore peggioramento, con coincidenza di poco e nulla. L'unico indice paragonabile è il terzo quartile, ma che nel caso dei SD coincide, addirittura, con l'estremo superiore. Il primo quartile dei SD si trova a livello dell'estremo inferiore dei dati originali e le mediane sono molto discoste. Anche gli outlier suggeriscono una scarsa affidabilità.

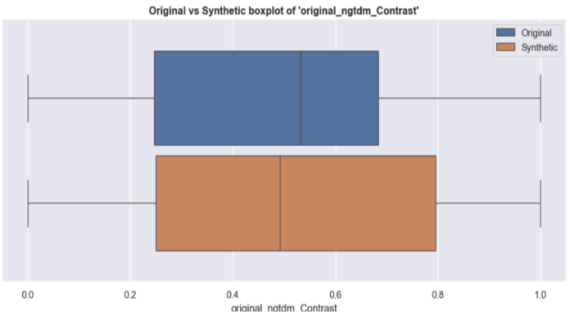

Figura 22 : Distribuzione feature "*original_ngdtm_Contrast*" del campione completo

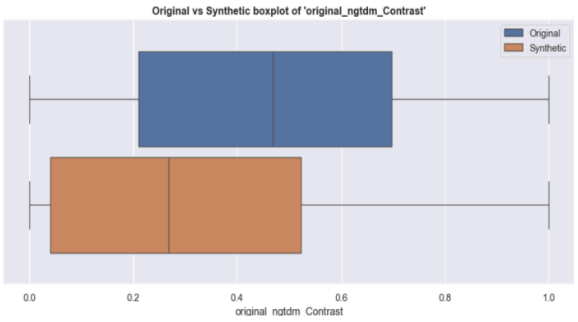

Figura 23 : Distribuzione feature "*original_ngdtm_Contrast*" del primo sottocampione (ADN)

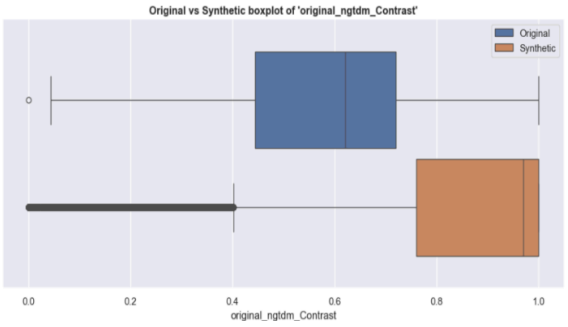

Figura 24: : Distribuzione feature "original_ngdtm_Contrast" del secondo sottocampione (SCC o NOS)

2.4. Risultati Quality Evaluation

Per ciascun metodo applicato (1.4.1, 1.4.2), successivamente, si è deciso di eseguire una quality evaluation. La quality evaluation è effettuata con 2 obiettivi: il primo era quello di osservare quanto erano fedeli i SD generati dai metodi a livello strutturale, cioè se non avevano valori mancanti e se i dati SD venivano riportati in una tabella come nei dati originali, che è stato fatto con un diagnostic report che valuta se i nomi delle features coincidono tra dati originali e SD e anche che i valori presenti nei dati originali e nei SD coincidano con il tipo di dato. Il secondo, è stato quello di misurare la similarità statistica tra dati originali e SD, che valuta la similarità delle

distribuzioni marginali di ogni feature e la similarità delle correlazioni tra le features dei dati originali e SD. Questi due metodi sono stati spiegati in dettaglio in (1.5).

Si inizia con un diagnostic report, in cui vengono valutate la data validity e la data structure. A prescindere dal numero di epoche impostato, tali metriche ottengono un punteggio complessivo del 100% per entrambi i metodi applicati, e anche per ciascun sottogruppo analizzato. Un punteggio, che ci aspettavamo dato che dal punto di vista strutturale ha una complessità non elevata dato che i dati sono tutti in una singola tabella e non hanno particolari vincoli che potrebbero portare ad una struttura più complessa.

Si procede quindi con il quality report rispetto alle metriche column shapes e columns pair trends. Guardando ai risultati in termini di overall score average tra quanto ottenuto per ciascuna metrica, abbiamo i seguenti punteggi:

	Tutta la popolazione	**Popolazione con ADN**	**Popolazione con SCC/NOS**
Metodo V&M	78%	79%	80%
Metodo CTGAN (300 epoche)	77%	75%	72%
Metodo CTGAN (600 epoche)	78%	77%	74%

Tabella 3 Risultati del Quality Report

Dai punteggi complessivi si nota che il metodo VM, per tutti e tre i sottogruppi considerati, ricalca meglio, rispetto al metodo del CTGAN fatto inizialmente con 300 epoche la forma e la distribuzione dei dati delle colonne reali in ogni colonna sintetica (columns shapes) e mantiene le relazioni e le dipendenze tra coppie di variabili osservate nei dati reali (column pair trends).

Si prova dunque un nuovo valore per il numero di epoch del metodo CTGAN, pari a 600, ricordando che troppe epoche possono portare a overfitting, dove il modello si adatta troppo ai dati di addestramento e non generalizza bene ai dati nuovi; ma d'altro canto poche epoch possono portare a underfitting, dove il modello non apprende abbastanza dai dati di addestramento. Nel nostro caso, i problemi legati

all'overfitting non si sono presentati dato che abbiamo pre-processato i dati rimuovendo le features altamente correlate per evitare ridondanza nell'addestramento del modello CTGAN.

Ciò che quindi risulta evidente è che, sicuramente, il CTGAN necessita di un maggior numero di cicli di addestramento per ottenere un punteggio migliore. Nonostante ciò, la quality evaluation evidenzia una maggior tendenza del metodo VM ad assicurare che i dati generati siano il più possibile indistinguibili dai dati reali.

2.5. Risultati degli errori dei metodi

Per analizzare i risultati, abbiamo calcolato la media degli errori relativi su tutte le feature per ottenere una misura complessiva della differenza tra i dati originali e quelli generati sinteticamente. La procedura è stata effettuata per ogni popolazione, sia per la popolazione completa, sia per la popolazione con presenza con ADN e la popolazione con SCC/NOS. Sono state calcolate le medie degli errori relativi degli indici statistici più importanti (media, varianza, asimmetria e curtosi). Nella Tabella 4 sono stati riportati ed evidenziati i risultati per ogni metodo.

		Mean	**Variance**	**Skewness**	**Kurtosis**
Whole tumorous population	**V&M**	75.29%	63.46%	3.20%	372.22%
	CTGAN (600)	31.23%	29.72%	134.05%	996.05%
Population with ADN	**V&M**	59.94%	63.24%	6.26%	23.92%
	CTGAN (600)	31.85%	39.30%	121.12%	141.90%
Population with SCC/NOS	**V&M**	34.73%	72.05%	8.31%	56.25%
	CTGAN (600)	33.07%	32.31%	884.08%	620.66%

Tabella 4 Errore relativo della media, la varianza, l'asimmetria e la curtosi per i due metodi VM e CTGAN.

In primo luogo, abbiamo calcolato l'errore relativo sulla media. Per la popolazione intera abbiamo ottenuto i seguenti valori: 75.29% per il metodo di VM e 31.23% per il metodo CTGAN. Per quanto riguarda la popolazione con adenocarcinoma (ADN), l'errore relativo è pari a 31.85% nel caso di CTGAN e 59,94% nel caso di VM. Nel caso invece della popolazione con cellule squamose (SCC o NOS) sono stati ottenuti valori molto simili con entrambi metodi: 34.73% per VM e 33.07% per CTGAN. Possiamo quindi evincere come i risultati ottenuti riguardanti la media con il metodo CTGAN riproducano in modo più fedele i dati reali.

In secondo luogo, abbiamo trattato l'errore relativo riguardante la varianza: per la popolazione intera sono stati ottenuti valori pari a 63,46% e 29.72%, per la popolazione con ADN pari a 63,24% e 39,30% e per la popolazione con SCC o NOS pari a 72,05% e 32,31% sempre rispettivamente nel caso VM e nel caso CTGAN.

In terzo luogo, abbiamo calcolato l'errore relativo relativamente all'asimmetria: per la popolazione intera sono stati ottenuti valori pari a 3,20% e 134,05%, per la popolazione con ADN pari a 6,26% e 121,12% e per la popolazione con SCC o NOS pari a 8,31% e 884,08% sempre rispettivamente nel caso VM e nel caso CTGAN. È possibile notare una grande discrepanza tra i valori, riscontrando come il metodo di VM fornisca risultati nettamente più simili ai dati reali.

Infine, abbiamo ricavato l'errore relativo della curtosi: per la popolazione intera sono stati ottenuti valori pari a 372,22% e 996,05%, per la popolazione con ADN pari a 23,92% e 141,90% e per la popolazione con SCC o NOS pari a 56,25% e 620,66% sempre rispettivamente nel caso VM e nel caso CTGAN. Abbiamo potuto riscontrare valori ottenuti con il metodo CTGAN piuttosto considerevoli.

Da ultimo, per confermare i risultati ottenuti attraverso la quality evaluation riguardo la similarità delle matrici di correlazione dei due metodi con la matrice di correlazione dei dati reali, abbiamo calcolato la norma di Frobenius per ogni matrice di "differenza" per poi metterle a confronto. Nel caso del metodo di VM abbiamo ottenuto un valore di 5.5640 per l'intera popolazione e 5.3993 e 6.0270 per la popolazione con presenza di ADN e quella con presenza di SCC/NOS rispettivamente. Mentre per il metodo CTGAN i valori della norma Frobenius sono stati maggiori per tutte le popolazioni rispetto al metodo di VM, con valori tra 20 e 21. Nella Tabella 5 sono riassunti i risultati della norma di Frobenius della differenza tra le matrici di correlazione, dove i valori tendenti a 0 avranno una maggior

similarità della matrice di correlazione, nello specifico nel metodo di VM avremo valori intorno a 5 che corrispondono a una similarità di circa il 90%.

	Whole population	ADN	SCC/NOS
Vale and Maurelli	5.5640	5.3993	6.0270
CTGAN (epoche=600)	20.1485	21.2807	21.2723
CTGAN (epoche=300)	20.2074	212904	21.2912

Tabella 5 Valori della norma di Frobenius della matrice delle differenze D per ogni metodo.

3 Conclusioni e Sviluppi Futuri

3.1. Discussione dei risultati

I risultati preliminari ottenuti dall'analisi esplorativa sono stati fondamentali per strutturare la generazione dei SD. In particolare, lo sbilanciamento tra le classi emerso dall'analisi ha portato a effettuare due generazioni distinte, una per il sottocampione che corrispondeva alla diagnosi di ADN e in cui erano contenute il 78% delle osservazioni del dataset e una per il sottocampione che corrispondeva a diagnosi di SCC o NOS. Questo approccio ha evitato che l'algoritmo favorisse una classe rispetto all'altra e ha permesso di rappresentare in modo più realistico le distribuzioni delle caratteristiche corrispondenti ai diversi valori di istologia. Si è osservata infatti, soprattutto nel sottocampione corrispondente all'istologia ADN, una maggiore somiglianza tra i SD e quelli reali rispetto a quando la generazione era stata effettuata sul campione completo.

Dall'analisi precedente, confrontando la distribuzione delle caratteristiche generate sinteticamente con quelle originali, si osserva che nel secondo sottocampione c'è una maggiore discrepanza tra i dati sintetici e quelli originali per entrambi i metodi utilizzati. Questo potrebbe essere dovuto al fatto che il secondo sottocampione include sia diagnosi di SCC che di NOS, risultando quindi meno omogeneo rispetto al primo sottocampione. Una minore omogeneità in un campione comporta una maggiore dispersione dei valori della sua distribuzione e una probabilità più alta di trovare outliers. Questi fattori influenzano negativamente la generazione dei dati sintetici. Se si effettua invece un confronto tra i metodi si ha in generale che, per quanto riguarda l'indice statistico della media, CTGAN è meno distante ai dati reali rispetto a VM, ma la differenza, comunque, non risulta estremamente sostanziale.

3.2. Conclusioni

I risultati della generazione dei SD per ogni metodo ci hanno fornito informazioni sufficienti per arrivare alle seguenti conclusioni. La prima cosa che si può evidenziare

è come il metodo CTGAN abbia una complessità di calcolo molto maggiore rispetto al metodo di VM nella generazione dei dati.

Questo viene spiegato dal fatto che il metodo del CTGAN, essendo un modello di DL, ha la necessità di essere allenato per poter generare dati di qualità, mentre il metodo di VM, essendo invece un metodo statistico, è più semplice dal punto di vista matematico e della implementazione. La differente complessità di calcolo rispecchia i tempi di elaborazione per la generazione dei dati: con il metodo VM la generazione dei dati ha richiesto un tempo considerevolmente inferiore (pochi secondi), rispetto al CTGAN (circa 5 minuti). In particolare si è visto come, al crescere delle iterazioni del CTGAN, il tempo computazionale sia cresciuto esponenzialmente (arrivando fino ad oltre 30 minuti).

La qualità dei SD generati valutata con la quality evaluation ha evidenziato come entrambi i metodi abbiano una qualità simile, ma con alcune differenze.

Ad esempio, la matrice di correlazione per il metodo di VM ha ottenuto una similarità maggiore alla matrice di correlazione dei dati originali rispetto alla matrice di correlazione del CTGAN. Mentre, il metodo del CTGAN all'aumentare del numero di iterazioni utilizzate per l'allenamento del modello, ha avuto una distribuzione marginale più simile alle distribuzioni delle features dei dati originali rispetto al metodo di VM.

Nell'analisi dell'errore relativo calcolato per i due metodi abbiamo sottolineato il fatto che l'errore relativo per la media e la varianza è stato minore con il metodo di CTGAN rispetto al metodo di VM corroborando i risultati del quality evaluation dove abbiamo visto come la distribuzione marginale fosse leggermente migliore rispetto a quella dei dati generati con VM. L'errore relativo dell'asimmetria con il metodo di VM è risultato molto minore rispetto all'errore relativo con il metodo CTGAN. È necessario prestare attenzione nell'utilizzo del metodo di VM , poiché ha mostrato delle limitazioni teoriche nella generazione di dati con determinate combinazioni di asimmetria e curtosi. La limitazione teorica del metodo di VM è stato il suo principale svantaggio quando messa a confronto con il metodo del CTGAN, dato che quest'ultimo genera dati tramite un generatore e un discriminatore che cercano di avvicinare il più possibile i valori generati sinteticamente a quelli veri senza limitarsi a determinati valori di curtosi o asimmetria.

Tuttavia, nel processo di generazione di SD con il CTGAN, questo presentava la problematica di overfitting, in cui i dati non vengono processati correttamente, il modello infatti è stato in grado di generare dati esattamente uguali a quelli originali, ma non risultava in grado di generalizzare i dati per osservazioni future.

Il vantaggio principale del metodo di VM, oltre alla sua semplicità di calcolo e basso costo computazionale, è stato la sua maggior similarità nelle correlazioni tra le features rispetto alle correlazioni reali. Questo è stato verificato nei nostri risultati dove nel quality report, la *Column Pair Trends* ha avuto un valore percentuale maggiore (circa il 95% per tutte due le popolazioni) rispetto a quello con il metodo del CTGAN (circa il 84% per tutte due le popolazioni). Risultato che è stato poi confermato dal calcolo della norma Frobenius, che indica la differenza tra la matrice di correlazione reale e quella generata con i metodi di generazione di SD. La matrice di correlazione calcolata sui dati generati con il metodo di VM è risultata essere notevolmente simile alla matrice di correlazione reale, confermando dunque i risultati della quality evaluation.

In sintesi, possiamo concludere elencando i vantaggi e gli svantaggi dell'utilizzo di ogni metodo per la generazione di SD.

Vantaggi del metodo CTGAN:

- E' stato in grado di preservare le relazioni pre-esistenti e la struttura delle distribuzioni dei dati originali
- Maggior similarità nelle distribuzioni marginali delle features reali

Svantaggi del metodo CTGAN:

- Maggior costo computazionale da 5 minuti per 600 iterazioni fino a 45 minuti per 1000 iterazioni
- Complessità elevata, poiché metodo DL
- Vulnerabilità all'overfitting e problemi nelle popolazioni con un elevato numero di features

Vantaggi del metodo di VM:

- Semplicità di calcolo
- Maggior similarità nelle correlazioni tra features

Svantaggi del metodo di VM:

- Limitazioni teoriche per alcune combinazioni di asimmetria e curtosi

3.3. Sviluppi futuri

In conclusione, una volta elencati i vantaggi e gli svantaggi di ogni metodo possiamo evincere la possibilità di individuare alcune differenze tra il metodo di VM e CTGAN per quanto riguarda la qualità dei dati generati. Migliorare l'architettura del modello e integrare metodi più robusti per la gestione dei valori mancanti e degli outliers potrebbe fare in modo che il metodo del CTGAN possa avere ottimi risultati; d'altra parte migliori tecniche di pre-processing dei dati reali e quindi una qualità maggiore dei dati in input, potrebbero generare in futuro risultati più realistici e un aumento della fedeltà al dataset originale; i dati fittizi assomiglieranno sempre di più ai dati reali.

Un'altra direzione su cui si potrebbe proseguire è quella di mettere a confronto altri metodi di generazione di SD o modificare gli iperparametri che si trovano all'interno del modello CTGAN.

Un'altra strada che si potrebbe percorrere è quella della creazione di linee guida dettagliate per la generazione di SD: risulta necessario sviluppare standard di conformità[4], ovvero protocolli in modo da regolare l'uso di SD per migliorare la ricerca collaborativa: avendo a disposizione dati standardizzati e linee guida precise infatti, risulterà più semplice la collaborazione tra istituti di ricerca, ospedali e aziende.

Vista l'alta qualità dei dati generati utilizzando i metodi sopracitati, si può dedurre che l'utilizzo di SD all'interno della radiomica sia un campo molto promettente, in grado di risolvere problematiche significative.

Bibliografia

[1] Mayerhoefer, M. E., Materka, A., Langs, G., Häggström, I., Szczypiński, P., Gibbs, P., & Cook, G. (2020). Introduction to radiomics. *Journal of Nuclear Medicine*, *61*(4), 488–495. https://doi.org/10.2967/JNUMED.118.222893

[2] Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I. F., Mak, R. H., Tamimi, R. M., Tempany, C. M., Swanton, C., Hoffmann, U., Schwartz, L. H., Gillies, R. J., Huang, R. Y., & Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, *69*(2), 127–157. https://doi.org/10.3322/caac.21552

[3] van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H., & Baessler, B. (2020). Radiomics in medical imaging—"how-to" guide and critical reflection. In *Insights into Imaging* (Vol. 11, Issue 1). Springer. https://doi.org/10.1186/s13244-020-00887-2

[4] Pascuzzo, R., Garattini, S. K., & Doniselli, F. M. (2024). Clinical Application of Radiomics in Oncology: Where Do We Stand? *Journal of Magnetic Resonance Imaging*. https://doi.org/10.1002/jmri.29340

[5] Jacobs, F., D'Amico, S., Benvenuti, C., Gaudio, M., Saltalamacchia, G., Miggiano, C., de Sanctis, R., della Porta, M. G., Santoro, A., & Zambelli, A. (2023). Opportunities and Challenges of Synthetic Data Generation in Oncology. *JCO Clinical Cancer Informatics*, *7*.

[6] Murtaza, H., Ahmed, M., Khan, N. F., Murtaza, G., Zafar, S., & Bano, A. (2023). Synthetic data generation: State of the art in health care domain. In *Computer Science Review* (Vol. 48). Elsevier Ireland Ltd. https://doi.org/10.1016/j.cosrev.2023.100546

[7] Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. In *Neurocomputing* (Vol. 493, pp. 28–45). Elsevier B.V.

https://doi.org/10.3390/cancers13123088

[8] Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-00977-1

[9] Achuthan, S., Chatterjee, R., Kotnala, S., Mohanty, A., Bhattacharya, S., Salgia, R., & Kulkarni, P. (2022). Leveraging deep learning algorithms for synthetic data generation to design and analyze biological networks. In *Journal of Biosciences* (Vol. 47, Issue 3). Springer. https://doi.org/10.1007/s12038-022-00278-3

[10] Endres, M., Mannarapotta Venugopal, A., & Tran, T. S. (2022). Synthetic Data Generation: A Comparative Study. *ACM International Conference Proceeding Series*, 94–102. https://doi.org/10.1145/3548785.3548793

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

[12] David, C., & Maurelli, V. A. (1983). SIMULATING MULTIVARIATE NONNORMAL DISTRIBUTIONS. In *NOTES AND COMMENTS* (Vol. 48, Issue 3).

[13] https://towardsdatascience.com/how-to-generate-real-world-synthetic-data-with-ctgan-af41b4d60fde

[14] Bakr, S., Gevaert, O., Echegaray, S., Ayers, K., Zhou, M., Shafiq, M., Zheng, H., Zhang, W., Leung, A., Kadoch, M., Shrager, J., Quon, A., Rubin, D., Plevritis, S., & Napel, S. (2017). Data for NSCLC Radiogenomics (Version 4) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2017.7hs46erv.

[15] https://www.gavazzeni.it/malattie/adenocarcinoma/

[16] https://www.skincancer.org/international/carcinoma-spino-cellulare/

[17] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic

Phenotype. Cancer Research, 77(21), e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

[18] https://pyradiomics.readthedocs.io/en/latest/features.html

[19] Abedi, M., Hempel, L., Sadeghi, S., & Kirsten, T. (2022). GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Applied Sciences (Switzerland)*, *12*(14). https://doi.org/10.3390/app1214707

[20] https://docs.sdv.dev/sdmetrics/metrics/metrics-glossary/correlationsimilarity

Lista delle Figure

Lista delle Tabelle