POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Biomedica

Tesi di Laurea Magistrale

Predizione rischio di recidiva e profilo mutazionale dei tumori stromali gastrointestinali tramite machine learning applicato ad immagini TC

Relatore

Prof.ssa Gabriella Balestra

Correlatori

Ing. Valentina Giannini

Ing. Samanta Rosati

Candidato

Giovanni Serra

Anno Accademico 2020-2021

Indice

1 Tumori stromali gastrointestinali – GIST

I tumori stromali gastrointestinali (GIST, GastroIntestinal Stromal Tumors) sono neoplasie rare che si possono sviluppare lungo tutto il tratto gastrointestinale, dall'esofago fino al retto.

Risultano essere meno dell'1% di tutte le neoplasie maligne, ma sono i tumori mesenchimali più comuni del tratto gastroenterico [1]. Essi hanno origine da cellule mesenchimali staminali del tratto gastrointestinale [2] e sono fenotipicamente correlati con le cellule interstiziali di Cajal (ICCs), delle cellule muscolari lisce del tratto gastrointestinale responsabili del movimento di contrazione peristaltica [3,4].

1.1 Prevalenza

In Italia, l'incidenza è di 1,5 casi per 100.000 persone ogni anno (tale stima tiene conto solamente dei GISTs clinicamente rilevanti, ossia maggiori di 1.5 cm in diametro) con una leggera prevalenza negli uomini adulti di età compresa tra i 60 e i 65 anni [5].

Le sedi di insorgenza più frequenti [1] sono:

- Stomaco (50%)
- Intestino tenue (25%)
- Retto (5%)
- Esofago (<5%)
- Localizzazioni extra-intestinali (<5%)

1.2 Patogenesi

I ricercatori hanno individuato nel 1998 la mutazione funzionale del KIT, confermando che circa l'80% dei GISTs presentava tale mutazione [6]. Le mutazioni a carico di KIT coinvolgono principalmente gli esoni 11 e 9 all'esordio della malattia, gli esoni 13, 14 e 17 in fase più avanzata; in rari casi, è stata documentata anche una mutazione a carico dell'esone 8 [2].

Nel 2003 è stata individuata un'ulteriore mutazione a carico del gene per il recettore PDGFR-α nel 5/10% dei casi analizzati. Tali mutazioni coinvolgono prevalentemente l'esone 18 e, meno

frequentemente, gli esoni 12 e 14 [2]. Le mutazioni di KIT e PDGFR- α sono mutualmente esclusive [2,6].

Tali geni codificano due recettori trans-membranali per la tirosin-chinasi. La chinasi presenta un ruolo determinante nella proliferazione cellulare, ed è considerata la "forza motrice" della patogenesi dei GISTs. Normalmente tali recettori sono attivati quando si legano alla tirosin-chinasi, ma in caso di mutazione essi possono auto-fosforilarsi senza bisogno di alcun ligando [3,4,6]; il risultato ultimo di questa attivazione è un aumento di proliferazione cellulare e un decremento dell'apoptosi, che porta in seguito alla formazione di una neoplasia generando successive mutazioni genetiche [7].

Negli altri casi, tra il 10 e il 15%, i GISTs non presentano mutazioni KIT o PDGFR- α; tali tipologie di tumori sono definite "Wild-Type", sono clinicamente indistinguibili dai precedenti, presentano un'identica morfologia, esprimono alti livelli di chinasi e non è ancora stato identificato il meccanismo di attivazione dei recettori [3,8]. Essi nella maggior parte dei casi si associano a forme sindromiche. I tumori WT possono essere suddivisi in SDH deficient e non-SDH deficient [5,6].

Il gruppo SDH deficient dal punto di vista clinico si associa spesso alla sindrome di Carney ed alla sindrome di Carney-Stratakis. Il gruppo non-SDH deficient è correlato ad alterazioni del gene NF-1, sviluppate solitamente in presenza di Neurofibromatosi di tipo 1; in rarissimi casi si osservano mutazioni dei geni BRAF e RAS [5,6].

1.3 Clinica

I GISTs sono comunemente asintomatici per dimensioni inferiori ai 6 cm di diametro.

In fasi più avanzate i GISTs cominciano ad assumere dimensioni maggiori determinando manifestazioni cliniche di diversa entità; esse possono essere meno gravi, come sazietà precoce, gonfiore e disfagia, o più gravi, quali sanguinamento addominale e anemia che può portare a una chirurgia d'emergenza [8].

Sintomi meno comuni includono nausea, dolore toracico pleurico, dolore pelvico e piccole ostruzioni intestinali [9].

Approssimativamente il 20% dei pazienti presenta metastasi al momento della diagnosi, localizzate principalmente nel fegato e nella cavità addominale [9]. Raramente i GISTs

metastatizzano nelle ossa, con una predilezione per la colonna vertebrale, e nei tessuti molli periferici; metastasi localizzate nei linfonodi e nei polmoni sono estremamente rare [4,10].

Nei GISTs aggressivi le metastasi si sviluppano entro i primi 2 anni, ma la loro diffusione può avvenire in un lasso di tempo molto lungo; sono stati osservati casi di diffusione dopo 42 anni [4]. Questo indica la necessità di un follow-up a lungo termine per i pazienti affetti da tale tumore.

1.4 Diagnosi

Data la loro natura asintomatica, la diagnosi è occasionale nel 21% dei casi [5, 11]; essi vengono spesso individuati in seguito ad un rilievo endoscopico, un rilievo radiologico o interventi chirurgici effettuati per altre patologie.

Nel 40% dei casi, la diagnosi avviene in urgenza per complicanze dovute al tumore quali perforazione o emorragia [5].

Talvolta, tali tumori vengono diagnosticati post-mortem in sede di autopsia [11].

Il GIST è considerato una neoplasia maligna, di conseguenza qualsiasi sospetto dovrebbe prevedere un accertamento istologico.

1.4.1 Diagnostica per immagini

Per la diagnosi di GIST sono necessari esami di diagnostica per immagini; le tecniche di imaging più utilizzate sono la tomografia computerizzata (TC), la risonanza magnetica (MRI) e la tomografia a emissione di positroni (PET) [5,9,15-18].

Una volta individuata un'area sospetta si deve necessariamente procedere con la biopsia, in quanto è necessario differenziare i GISTs da altre neoplasie addominali come linfomi, neoplasie germinali ed altri tipi di sarcomi. [5]

1.4.2 Diagnostica istologica

Il materiale istologico può essere ottenuto tramite atto chirurgico, tecniche endoscopiche o biopsie transcutanee.

I GISTs sono solitamente diagnosticati tramite immunistochimica per individuare la presenza della proteina KIT, presente nel 90% dei casi (per tumori derivanti da mutazioni KIT o PDGFR-α) ma non negli altri tumori gastrointestinali.

Nel caso dei GISTs Wild Type si ricercano mutazioni della proteina NF1 (in caso di neurofibromatosi di tipo 1) o del RAS/BRAF. [9]

La diagnosi istopatologica deve includere anche l'indice mitotico, la sede di origine del tumore, le sue dimensioni e l'integrità della capsula tumorale. [5]

1.5 Stratificazione del rischio di recidiva

Sono stati individuati quattro fattori prognostici per i GISTs:

- Dimensioni del tumore; il valore di cutoff è 5 cm. I tumori più piccoli del cutoff presentano una prognosi migliore
- Indice mitotico; i GISTs presentano tipicamente un indice mitotico basso, per questo motivo viene valutato il numero di mitosi per 50 High Power Fields (HPFs) contro i classici 10 HPFs.
- Sede di origine del tumore
- Rottura della capsula tumorale, che determinerebbe un aumento significativo del rischio di caduta

Questi fattori condizionano il comportamento biologico della malattia e sono utili a determinare la prognosi del paziente. [1,5,9]

L'analisi mutazionale KIT e PDGFR-α risulta essere un fattore prognostico e predittivo di risposta a terapia farmacologica, ma nonostante questo non è ancora stata inserita in nessuna classificazione del rischio, in quanto non correlata significativamente con il rischio di recidiva. Essa viene comunque indicata in quanto predittiva della risposta al trattamento farmacologico [5].

Negli anni sono stati proposti diversi sistemi per la stratificazione del rischio di recidiva basati su questi fattori; la valutazione del rischio è fondamentale per stabilire la terapia da adottare per la cura del tumore e decidere la modalità di esecuzione del follow up, anche se ad oggi non esistono protocolli basati su prove di efficacia [5].

Analizziamo tali sistemi nello specifico.

1.5.1 National Institutes of Health (NIH) Consensus Criteria

Il primo sistema di stratificazione del rischio, sviluppato da Fletcher et al. nel 2002; l'NIH si basa sulla dimensione del tumore e sull'indice mitotico per individuare quattro classi di rischio: molto basso, basso, intermedio e alto [12]. Nella Tabella 1 è possibile vedere i criteri di stratificazione:

Tabella 1. *NIH Consensus Criteria*

Risk category	Tumor size (cm)	Mitotic count (per 50 HPF)
Very low	<2	<5
Low	2-5	<5
Intermediate	<5	6-10
	5-10	<5
High	>5	>5
	>10	Any
	Any	>10

1.5.2 Armed Forces Institute of Pathology Criteria

L'AFIP, anche detto Miettinen's Criteria [1,12], a differenza dell'NIH Consensus include tra i fattori prognostici anche la sede originaria del tumore. Tale inclusione risulta essere molto utile, in quanto a parità di dimensioni ed indice mitotico i GISTs gastrici hanno una prognosi migliore rispetto ai GISTs del piccolo intestino [9,10].

Questo criterio suddivide la dimensione dei tumori in 4 gruppi, la conta mitotica in due e differenzia le varie sedi originarie del tumore; queste tre variabili contribuiscono a classificare i tumori in 8 sotto-gruppi (1-6b) che corrispondono a 5 classi di rischio: nessuno, molto basso, basso, moderato ed alto, come visibile in Tabella 2.

A differenza dell'NIH Consensus Criteria, l'AFIP introduce la classificazione dei GISTs benigni. [12]

Tabella 2. *AFIP Criteria*

Group	Tumor size (cm)	Mitotic count (per 50 HPF)	Stomach	Small intestine	Duodenum	Rectum
1	≤2	≤5/50	None	None	None	None
2	>2 to ≤5		Very Low	Low	Low	Low
3a	>5 to ≤10		Low	Moderate	Insufficient data	Insufficient data
3b	>10		Moderate	High	High	High
4	≤2	>5/50	None	High	Insufficient Data	High
5	>2 to ≤5		Moderate	High	High	High
6a	>5 to ≤10		High	High	Insufficient data	Insufficient data
6b	>10		High	High	High	High

1.5.3 Modified NIH Criteria (Joensuu Risk Criteria)

L'NIH originale proposto da Fletcher et al. nel 2002 presentava diverse limitazioni: non incorporava tra i fattori prognostici la sede del tumore, non teneva conto della rottura della capsula tumorale e non definiva la classificazione del rischio per i GISTs che avevano esattamente 5 mitosi per 50 HPF.

Per ovviare a questi limiti, Joensuu propose una versiona modificata di tale sistema, includendo i fattori prognostici mancanti e stabilendo la rottura della capsula tumorale come fattore di rischio assoluto, fornendo i seguenti criteri visibili in Tabella 3. [12]

Tabella 3*. Modified NIH Criteria*

Risk category	Tumor size (cm)	Mitotic count (per 50 HPF)	Primary tumor site
Very low	<2.0	≤5	Any
Low	2.1-5.0	≤5	Any
Intermediate	2.1-5.0	>5	Gastric
	<5.0	6-10	Any
	5.1-10.0	≤5	Gastric
High	Any	Any	Tumor rupture
	>10	Any	Any
	Any	>10	Any
	>5.0	>5	Any
	2.1-5.0	>5	Non-gastric
	5.1-10.0	≤5	Non-gastric

1.6 Terapia

Secondo le linee guida NCCN (National Comprehensive Cancer Network) i GISTs possono essere distinti in quattro differenti categorie, ognuna con le proprie strategie terapeutiche [13]:

- GISTs gastrici molto piccoli
- GISTs localizzati o potenzialmente asportabili
- GISTs non asportabili o metastatici

Prima di cominciare la terapia, i GISTs devono esser valutati tramite CT o MRI; data la possibilità che i GISTs metastatizzino alle ossa, è consigliata una PET in caso di comparsa di dolori ossei o aumenti di fosfatasi alcalina.

Altro esame necessario prima di iniziare la terapia è quello dell'analisi mutazionale, necessario per predire l'efficacia di un eventuale terapia adiuvante farmacologica.

L'imatinib mesilato è un farmaco antitumorale mirato sviluppato sulla base dell'eziologia dei GISTs; tale farmaco era già utilizzato come trattamento per la leucemia mieloide cronica ma, in seguito, è stato scoperto che inibiva l'attività della tirosina chinasi inibendo del tutto il legame ATP di KIT e PDGFR-α.

Nei GISTs abbiamo un'attivazione costante del recettore per la tirosina chinasi e, tramite l'utilizzo di questo farmaco, si è in grado di inibire tale attivazione rallentando il decorso del tumore e, in alcuni casi, regredendoli.

La funzionalità del farmaco dipende dalla mutazione che ha dato origine al GIST:

- Mutazione KIT sull'esone 11 sensibile all'imatinib con dose standard di 400mm/gg
- Mutazione KIT sull'esone 9 sensibile all'imatinib con dose più alta rispetto allo standard
- Mutazione PDGFR-α sull'esone 18 scarsissima sensibilità all'Imatinib
- GISTs Wild Type non rispondono alla terapia con imatinib

Per questo motivo è necessario fare un'analisi mutazionale prima di procedere alla terapia; a seconda della mutazione, si procederà con la somministrazione della terapia farmacologica adiuvante o si procederà in altro modo.

Dopo circa due anni, la metà dei pazienti sottoposti a terapia con imatinib ha sviluppato una resistenza al farmaco dovuta ad una mutazione secondaria del gene che previene l'accoppiamento tra il recettore per la tirosina chinasi e l'imatinib, annullandone del tutto l'efficacia.

Vediamo le strategie di intervento applicate sulle varie tipologie di GISTs esistenti, che si suddividono in GISTs asportabili (very small o localizzati) e non asportabili. [13]

1.6.1 Very small GISTs

I GISTs appartenenti a questa categoria possono essere individuati durante esami endoscopici o CT; in tal caso, una diagnosi patologica dovrebbe essere effettuata utilizzando la tecnica dell'agoaspirazione sotto guida eco-endoscopica (EUS-FNA) [14].

La strategia per i GISTs asportabili segue il protocollo mostrato in Figura 1.

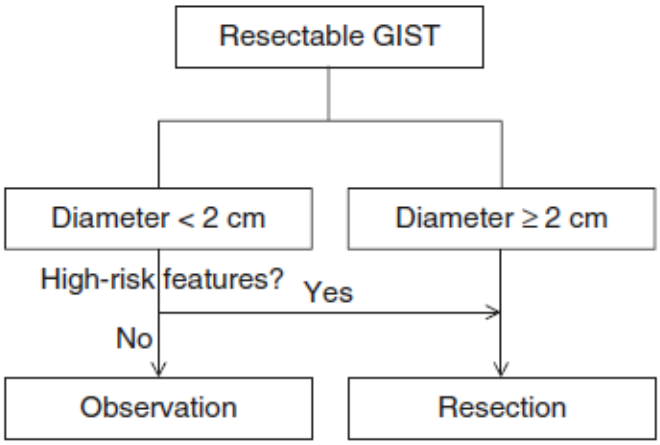

Figura 1. *Strategia di intervento per GISTs asportabili*

In caso di assenza di fattori di alto rischio come bordi irregolari, ulcerazioni e aree ecogeniche, è raccomandata una sorveglianza periodica tramite radiografia.

Una sorveglianza periodica tramite EUS-FNA è da tenere in considerazione solo dopo aver discusso con il paziente di rischi e benefici di tale tecnica nel suo specifico caso.

In presenza di fattori ad alto rischio, si procede invece con la resezione chirurgica; in seguito, se il rischio di recidiva è basso, si procede con una sorveglianza periodica. In caso di alto rischio, invece, si segue con somministrazione di imatinib per almeno 36 mesi con annessa sorveglianza periodica ogni 6/12 mesi per cinque anni [13]

1.6.2 GISTs localizzati

Essi si suddividono in due ulteriori categorie [13]:

- Inizialmente asportabili. L'asportazione chirurgica di questi tumori è sempre consigliata; nel caso di localizzazione anatomica appropriata, si può procede anche con asportazione laparoscopica.

  È fondamentale per il chirurgo porre estrema attenzione a preservare l'integrità della capsula tumorale al fine di diminuire notevolmente il rischio di recidiva.

  È necessario determinare il rischio di recidiva post-intervento per stabilire il percorso da seguire; in caso di alto rischio si procede con la somministrazione di imatinib per almeno 36 mesi dopo l'intervento unito a sorveglianza periodica ogni 3/6 mesi per 5 anni e imaging annuale.

  Se il rischio è basso, non è consigliata la somministrazione di imatinib ma solamente una sorveglianza periodica ogni 3/6 mesi con annesso imaging annuale.

  Se l'intervento non dovesse essere andato a buon fine e vi fosse un importante residuo tumorale, è fortemente consigliata l'assunzione di imatinib fino alla scomparsa del tumore, comparsa di eventi avversi o evoluzione della malattia; un altro intervento non è indicato.
- Inizialmente non asportabili. Talvolta i GISTs localizzati possono essere difficili da asportare in quanto difficilmente raggiungibili; in questi casi una terapia adiuvante preoperatoria con imatinib può essere utile per ridurre il tumore e rendere più facile l'asportazione.

  In questi casi è necessaria una stretta sorveglianza del paziente al fine di tenere sotto controllo eventuali effetti avversi o una rapida crescita tumorale.

1.6.3 GISTs metastatici non asportabili

Per quanto riguarda i GISTS non asportabili, la strategia di intervento è più strutturata rispetto alla precedente, come è possibile notare in Figura 2 [13].

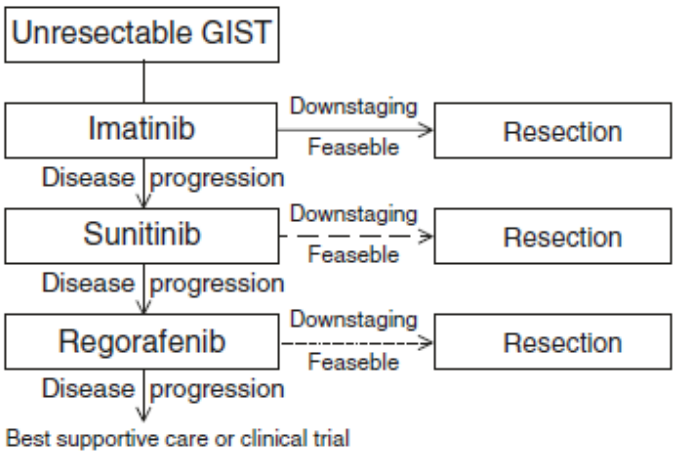

Figura 2. *Strategia di intervento per GISTs non asportabili*

In questi casi vi sono varie fasi di trattamento:

- Prima fase: analisi mutazionale per verificare se il GIST è sensibile al trattamento con imatinib. In tal caso si procede con la somministrazione e, se le dimensioni del tumore diminuiscono, si procede con la chirurgia per asportare il tumore. Se la chirurgia non fosse possibile, si proseguirebbe con la somministrazione di imatinib fin quando non si presentano eventi avversi o evoluzioni della malattia
- Seconda fase: se si nota un progresso della malattia tramite imaging, si consiglia la sostituzione dell'imatinib con il sunitinib, farmaco più efficace per i tumori con mutazioni wild type
- Terza fase: nei casi in cui la somministrazione di imatinib e sunitinib non abbia portato alcun effetto benefico, si prosegue il trattamento con la somministrazione di regorafenib

Il regorafenib risulta essere l'ultima opzione farmacologica al momento disponibile, se anch'essa dovesse fallire si proporrebbe al paziente la partecipazione a trial clinici [13].

2 Machine Learning e immagini TC

2.1 Obiettivi del lavoro

La stratificazione del rischio di recidiva unita all'analisi mutazionale sono due parametri necessari per predire l'efficacia di un'eventuale terapia adiuvante farmacologica e per stabilire un piano d'azione ottimale per la cura dei pazienti affetti da tumore stromale gastrointestinale (GIST).

L'obiettivo di questo lavoro di tesi è quello di costruire due classificatori che siano in grado di ottenere tali parametri partendo dalle sole features radiomiche delle immagini TC, evitando così di dover sottoporre il paziente a biopsia.

Per quanto riguarda il classificatore del rischio, sono state individuate due macro-classi che racchiudessero diverse categorie individuate tramite il sistema di stratificazione del rischio AFIP:

- Prognosi favorevole (classe 0), pazienti con rischio None, Very low e low
- Prognosi Sfavorevole (classe 1), pazienti con rischio Intermediate, High e i pazienti metastatici.

Nel caso del classificatore dell'analisi mutazionale:

- Mutazione KIT (classe 0), pazienti che rispondono bene a terapia adiuvante imatinib
- Mutazione PDGFR-α e Wild Type (classe 1), pazienti non responsivi all'imatinib

2.2 Materiali e metodi

2.2.1 Popolazione dello studio

Sono stati coinvolti nello studio 82 pazienti affetti da GIST, 52 dei quali afferenti presso l'Istituto di Candiolo IRCCS (TO), i restanti 30 presso l'A.O.U. Policlinico Paolo Giaccone (PA). Tutti dovevano rispettare i seguenti criteri:

- Diagnosi di GIST confermata da analisi istopatologica con annesso profilo mutazionale e conta mitotica

- Acquisizione dell'immagine TC con mezzo di contrasto in fase venosa portale pretrattamento farmacologico adiuvante
- Lesioni con diametro minimo di 1,5 cm al fine di ottenere una Region Of Interest appropriata
- Nessun trattamento biologico o chemioterapico svolto nei due mesi precedenti all'acquisizione dell'immagine TC
- Assenza di altri tumori

2.2.2 Analisi istologica-genetica ed esami strumentali

Per tutti i pazienti arruolati nello studio sono stati forniti:

- Esame TC addome/torace baseline con mezzo di contrasto in formato DICOM.
- Maschere di segmentazione tumorali delle immagini TC. Tali immagini sono state segmentate manualmente da parte di un medico specializzando in radiodiagnostica includendo due strati: uno per la lesione in tutta la sua interezza, uno contenente soltanto le aree necrotiche. Grazie a questa differenziazione, è stato possibile escludere le aree necrotiche dallo studio.
- Analisi mutazionale del genotipo su reperto istologico proveniente dal tumore primitivo.
- Stratificazione del rischio di recidiva secondo criterio AFIP.

2.2.3 Dataset iniziale

Analizzando il dataset completo, sono state riportate in Tabella 4 le distribuzioni dei pazienti nelle varie classi.

Tabella 4. *Suddivisione pazienti dataset completo. *Dati sul rischio mancanti per sei pazienti Candiolo*

	Candiolo IRCCS (TO)	A.O.U. Giaccone (PA)	Tot.
# Pazienti	52	30	82
Mutazione			
0	33	22	55
1	19	8	27
Rischio Miettinen*			
0	24	10	34
1	22	20	42

Come è possibile notare, il dataset risulta essere sbilanciato dal punto di vista mutazionale; tale fatto era prevedibile, dato che le mutazioni KIT sono molto più comuni rispetto alle mutazioni PDGFR-α e Wild Type.

Relativamente al rischio AFIP, invece, si può osservare una distribuzione bilanciata tra le due classi.

La presenza di pazienti afferenti da due centri differenti offre la possibilità di suddividere il dataset in construction set (composto da pazienti di Candiolo) e validation set esterno (pazienti dell'A.O.U. Giaccone), per fornire maggior significatività allo studio in questione.

2.2.4 Pre-processing

Prima di cominciare la fase di costruzione dei classificatori, si è deciso di svolgere una piccola fase di pre-processing.

Tra gli 82 pazienti costituenti il dataset, 8 presentavano due maschere tumorali segmentate da due diversi operatori. La segmentazione tumorale è fatta manualmente, tramite l'ausilio di un software, da un operatore umano; tale fatto può determinare una certa variabilità della segmentazione inter-operatore, che porta inevitabilmente ad una variabilità delle features radiomiche con le quali andremo ad allenare i nostri classificatori. Se tali features dovessero variare notevolmente a seconda dell'operatore di competenza, il classificatore potrebbe risentirne ottenendo performance scadenti su immagini non utilizzate per il suo training.

Per questo motivo, si è deciso di sfruttare la presenza della doppia maschera tumorale al fine di valutare la stabilità delle features estratte, rimuovendo dal dataset tutte le features che presentavano una variabilità troppo elevata per permettere al classificatore di ottenere delle buone performances.

Per ogni paziente sono state prese in esame le due maschere tumorali, rimuovendo le zone necrotiche come specificato in precedenza; ogni immagine TC è stata normalizzata tra il primo e il 99esimo percentile (per evitare una normalizzazione su dei valori outliers) per permettere un confronto più significativo tra immagini provenienti da macchinari differenti.

Sono state estratte le features delle immagini TC sfruttando il "RadiomicsFeatureExtractor", un'oggetto facente parte della libreria "pyRadiomics" di Python [19]; sono stati impostati i seguenti parametri di estrazione:

- binCount = 32

- force2D = True. Essendo il voxel non isotropico, si è deciso di evitare un'interpolazione che avrebbe aggiunto informazione finta.
- Normalize = True

Prendendo in input l'immagine TC e la relativa maschera, esso è in grado di estrapolare tutte le features radiomiche utilizzate per questo studio, nello specifico 107 così suddivise:

- Shape: 14
- FirstOrder: 18
- GLCM: 24
- GLDM: 14
- GLRLM: 16
- GLSZM: 16
- NGTDM: 5

Tale processo è stato ripetuto due volte per gli 8 pazienti aventi la doppia segmentazione, prendendo ad ogni iterazione una maschera diversa; in questo modo son stati ottenuti due dataset aventi dimensione 8x107.

Per determinare la stabilità di tali features è stata calcolata per ognuna di esse la variazione percentuale con l'Equazione 1:

$$Variazione\ percentuale = \frac{feature_{maschera2} - feature_{maschera1}}{feature_{maschera1}} * 100$$

***Equazione 1**. Variazione percentuale features doppia segmentazione*

Per tutte le features degli 8 pazienti abbiamo così ottenuto una variazione percentuale; è stata fatta la media di tali variazioni sui diversi pazienti, ottenendo una variazione media percentuale delle 107 features a nostra disposizione.

Alla variazione media percentuale è stata affiancata anche una deviazione standard, per verificare se tale variazione differisse da paziente a paziente o risultasse stabile in tutti i casi.

Entrambe le distribuzioni così ottenute sono state graficate tramite istogramma per visualizzarne l'andamento e impostare un cutoff che permettesse l'esclusione dall'analisi delle features più instabili; l'istogramma della variazione percentuale media è visibile in Figura 3, quello della deviazione standard in Figura 4.

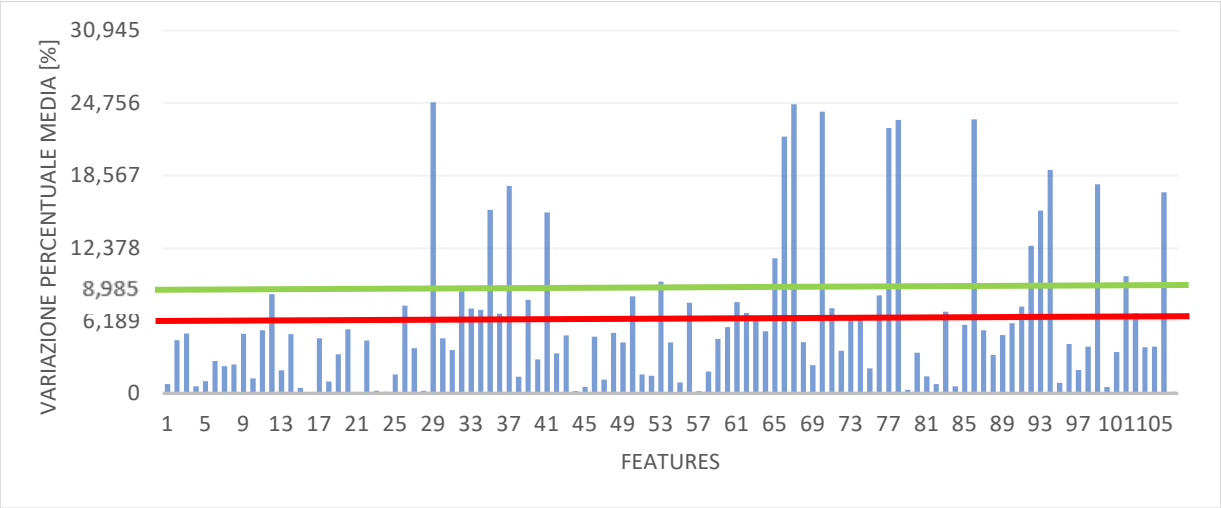

Figura 3. *Istogramma variazione percentuale media features*

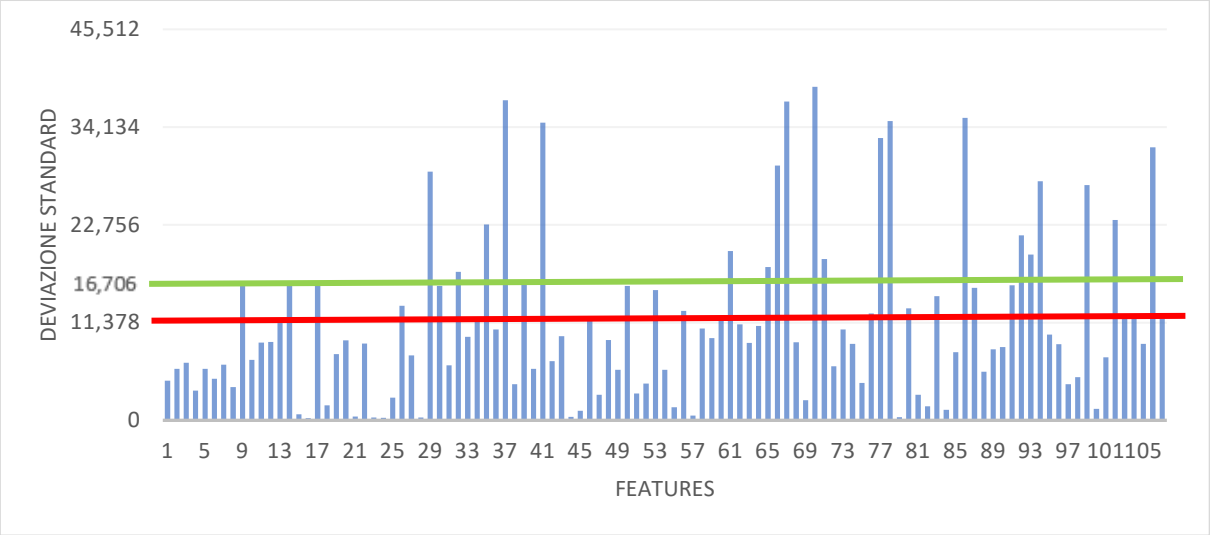

Figura 4. *Istogramma deviazione standard features*

Nelle Figure 3 e 4 sono riportati due cutoff: in rosso vediamo il valore medio delle distribuzioni, in verde il cutoff impostato per individuare quali features son da considerarsi stabili e quali no. Tale cutoff definitivo è stato stabilito visivamente, andando ad includere anche quelle features che si avvicinavano alla variazione percentuale media ma che sarebbero state scartate per una differenza poco significativa.

Sono state rimosse dallo studio tutte le features che presentavano:

- Variazione percentuale media > 8,985 %

- Deviazione standard > 16,706 %

Si è così ridotto il numero delle features radiomiche da 107 ad 86, osservando le seguenti variazioni:

- Shape: 14 → 14
- FirstOrder: 18 → 16
- GLCM: 24 → 20
- GLDM: 14 → 9
- GLRLM: 16 → 12
- GLSZM: 16 → 11
- NGTDM: 5 → 4

Grazie a questo lavoro di pre-processing, quindi, sono state rimosse 21 features instabili che avrebbero potuto influenzare negativamente le capacità dei classificatori che saranno successivamente costruiti.

2.2.5 Costruzione dataset di partenza

Una volta individuate le features utili per questo studio, sono state date in input al Features Extractor le immagini TC con la relativa maschera di tutti e 82 i pazienti arruolati, ottenendo il dataset completo avente dimensioni 82x87.

Per ogni paziente sono stati forniti anche la stratificazione del rischio AFIP e la mutazione genetica, che sono state aggiunte al dataset come label arrivando alle dimensioni finali di 82x89.

Come già detto in precedenza, i pazienti arruolati per questo studio provengono da due strutture differenti; questo fatto può essere sfruttato per effettuare uno studio multicentrico, suddividendo il dataset in construction e validation set. In questo modo, dopo aver costruito e allenato i classificatori, potremo validarli su pazienti esterni completamente estranei al processo di creazione del classificatore, rendendo lo studio più significativo. Sono quindi stati ottenuti i seguenti set:

- Construction set, costruito utilizzando i pazienti afferenti a Candiolo. Dimensioni 52x89
- Validation set, costruito con i pazienti dell'A.O.U. Giaccone. Dimensioni 30x89

2.2.6 Feature selection e costruzione classificatori

Per costruire i due classificatori desiderati è stato stabilito un "workflow" tipo che comprendeva i seguenti passaggi:

- Feature Selection
- Tuning dei parametri del classificatore scelto
- Allenamento del classificatore
- Validazione su validation set

Per quanto riguarda la feature selection, son stati testati e confrontati diversi metodi per verificare quale fornisse migliori risultati; tra i metodi provati troviamo:

- Nessuna feature selection, verranno utilizzate tutte le features a disposizione.
- Correlazione di Pearson; viene calcolata la correlazione delle singole features con la label di interesse. Tutte le features aventi p-value < 0.05 saranno selezionate, le altre verranno scartate.
- One Way ANOVA (ANalysis Of VAriances) test, calcola l'utilità delle singole features e le ordina dalla più utile alla meno utile. Quando si utilizza questo metodo di feature selection, verranno testate tutte le features in ordine di utilità, sommando ad ogni iterazione la feature successiva nell'elenco per allenare il classificatore; confrontando poi i classificatori allenati con i vari subset di features così selezionate, si sceglie la combinazione di features che ha contribuito all'ottenimento dei risultati migliori
- Least Absolute Shrinkage and Selection Operator (LASSO), è un embedded method che tiene conto anche dei subsets delle features a nostra disposizione; esso svolge un processo di "shrinking" che penalizza i coefficienti delle variabili di regressione, fino ad azzerarli. Le variabili che hanno coefficienti diversi da zero dopo lo shrinking process sono selezionate per far parte del modello.
- ANOVA + LASSO, combinazione dei due metodi precedentemente presentati che prende in considerazione le features selezionate contemporaneamente da entrambi i metodi.
- Importance calcolata tramite Random Forest; l'importance delle singole features viene calcolata in base a quanto diminuisce l'entropia in un albero. Anche in questo caso, come per l'ANOVA, verranno testati tutti i subset in ordine di importance.

Per ognuno di questi metodi di Feature selection, saranno valutate diverse tipologie di classificatori al fine di trovare la miglior combinazione possibile per i dati di questo studio; tra i classificatori testati vi sono:

- Bayesiano, è un classificatore con approccio probabilistico che si basa sul Teorema di Bayes; tale teorema descrive la probabilità di un evento sfruttando la conoscenza preliminare delle condizioni che potrebbero essere correlate all'evento stesso (in questo caso, le features). Esso calcola la probabilità che un evento A (classe di appartenenza) si verifichi, dato che B si è verificato (feature). In presenza di più features deve esser valida l'ipotesi di indipendenza, ossia ogni feature dev'essere indipendente l'una dall'altra; in questo modo è possibile suddividere prove multiple e trattarle come indipendenti, semplificando di gran lunga il calcolo delle probabilità.

  Il bayesiano è un algoritmo supervisionato utile per risolvere problemi di classificazione; esso sfrutta i dati facenti parte del training set per calcolare la probabilità condizionata di appartenenza ad una classe date tutte le features a sua disposizione. Più il dataset di allenamento è corposo, più saranno significative le probabilità calcolate.

  È un classificatore facile da implementare e veloce da allenare/testare, ma d'altra parte non è facile stabilire se le features siano effettivamente indipendenti tra di loro e, nella pratica, esistono sempre delle dipendenze tra alcune variabili che non possono essere modellate dal classificatore [20].
- SVM, è un classificato binario lineare non probabilistico che può essere utilizzato per scopi di classificazione o regressione. Questo modello cerca di stabilire una buona separazione degli elementi appartenenti al training, cercando di ottimizzare un confine lineare tra di essi; se i dati sono multidimensionali, il modello cercherà di creare un iperpiano che sia in grado di separare gli elementi appartenenti alle differenti classi. Si tratta di risolvere un problema di ottimizzazione con un criterio che ottimizzi la separazione, massimizzando la distanza intercorsa tra elementi delle due diverse classi in modo da ridurre al minimo i casi di errata classificazione.

  Questo classificatore funziona bene anche se allenato con dataset aventi pochi elementi, anche se è necessario che essi siano rappresentativi delle rispettive classi; solitamente nei training set sono contenuti anche elementi di "rumore" che aggiungono dell'imprevedibilità alla fase di training, rendendo più difficile la separazione.

  In alcuni casi non è possibile separare linearmente gli elementi di due classi; in queste situazioni si passa ad uno spazio dimensionale maggiore, aggiungendo dimensioni artificiali matematicamente dipendenti dall'originale. In questi piani, i dati possono

essere facilmente separabili. Il problema è che questo processo risulta difficile da implementare nella pratica, in quanto si dovrebbe aggiungere una dimensione "fittizia" e risolvere un problema di ottimizzazione nel nuovo spazio, rendendo questo processo computazionalmente pesante. Per evitare questi problemi si sfrutta il "kernel trick", che permette di risolvere il problema di ottimizzazione nello spazio originale [21].

- k-Nearest Neighbor (kNN), è un algoritmo supervisionato che individua i pattern per la classificazione di oggetti basandosi sulle features degli oggetti vicini ad esso. Solitamente la somiglianza viene calcolata tramite la distanza (che può essere calcolata in vari modi) che intercorre tra l'elemento da classificare e gli elementi del training set, minore sarà quest'ultima e maggiore sarà la somiglianza tra l'elemento da classificare e l'elemento appartenente al training set. Il kNN prevede un parametro k che fissa il numero di data points più vicini da analizzare; la classe che presenta il maggior numero di elementi più vicini all'elemento da classificare sarà scelta come previsione [22].
- Random Forest, è un algoritmo supervisionato che sfrutta i decision trees come modello individuale; combinando svariati decision trees in un unico modello, si costruisce una random forest. Se prese singolarmente, le previsioni fatte da questi alberi potrebbero risultare poco accurate, ma combinate insieme saranno decisamente più tendenti al risultato corretto. Il risultato finale sarà dato dalla classe restituita dal maggior numero di decision trees [23].
- XGBoost, un algoritmo di tipo "Gradient Boosting" simile ad un Random Forest, dove ogni albero è allenato sull'errore precedente [24].

2.2.7 Tuning parametri classificatori

Ogni classificatore presenta i suoi parametri caratteristici, al variare dei quali cambia notevolmente la capacità del classificatore stesso di performare con dati diversi a seconda dei casi; è stato svolto un processo di tuning di tali parametri per ogni classificatore per individuare il setting ideale per il nostro studio.

In Tabella 5 vediamo quali parametri son stati presi in considerazione.

Tabella 5. *Parametri classificatori*

Classificatore	Parametri
Bayesiano	/
SVM	C: 1,10,50,100
	Kernel: linear, poly, rbf, sigmoid
	Degree: 2,3
	Gamma: scale, auto
	Shrinking: True, False
	Decision_function_shape: ovo, ovr
	Probability: True, False
kNN	N_neighbors: 1,3,5,7,9,11,13,15,17,19
	Weights: uniform, distance
	Algorithm: auto, ball_tree, kd_tree, brute
	P: 1,2
	Metric: Euclidean, Manhattan, Chebyshev, Minkowski
Random Forest	N_estimators: 10,50,75,100,125,150,200
	Criterion: gini, entropy
	Max_features: auto, sqrt, log2
	Bootstrap: True, False
XGBoost	Learning_rate: 0.15, 0.1, 0.05, 0.01, 0.005, 0.001
	N_estimators: 100, 250, 500, 750, 1000, 1250, 1500, 1750
	Max_depth: 2,3,4,5,6,7
	Min_samples_split: 2,4,6,8,10,20,40,60,100
	Min_samples_leaf: 1,3,5,7,9
	Max_features: auto, sqrt, log2, None
	Subsample: 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1

Per fare il tuning dei parametri ci sono due strade percorribili:

- GridSearch, viene provata ogni combinazione possibile per poi estrarre il setting di parametri più performante. È computazionalmente costoso ma migliore in termini di risultati. È il metodo scelto per tutti i classificatori ad eccezione dell'XGBoost, che risulta essere molto più dispendioso in termini di potenza e tempo necessari per provare tutte le combinazioni.
- RandomSearch, vengono provate diverse combinazioni di parametri random per un numero fissato di iterazioni, cercando la combinazione migliore.

Per capire quale è la combinazione migliore ci si affida a dei parametri di scoring; il setting che massimizza tale parametro sarà quello considerato migliore. Per questo studio ne son stati scelti due:

- Accuracy per il classificatore del rischio; l'accuracy è la percentuale di predizioni che il modello è in grado di classificare correttamente rispetto a tutte le predizioni effettuate. Essa si calcola come nell'Equazione 2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equazione 2. *Formula Accuracy*

Dove TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives

- Balanced Accuracy per il classificatore della mutazione; a differenza della metrica sopra descritta, in questo caso si tiene conto dei datasets fortemente sbilanciati. Essa è ottenuta facendo la media aritmetica tra la sensitivity (percentuale di positivi correttamente identificati) e la specificity (percentuale di negativi correttamente identificati), come espresso nell'Equazione 3.

$$Balanced\ Accuracy = \frac{\left(\frac{TP}{TP + FN}\right) + \left(\frac{TN}{TN + FP}\right)}{2}$$

Equazione 3. *Formula Balanced Accuracy*

Tale differenza è dovuta al fatto che i dataset nei due casi sono diversamente bilanciati. Nel caso del rischio il construction set è bilanciato, di conseguenza l'accuracy è una buona metrica per valutare le prestazioni del classificatore; nel caso della mutazione, invece, il construction è fortemente sbilanciato a favore del gene KIT. Se scegliessimo come metrica di scoring l'accuracy, otterremmo dei modelli con un'elevata accuracy ma con pessime performance, in quanto potrebbe classificare bene le mutazioni KIT e non riconoscere l'altra classe, pur mantenendo un'elevata accuracy. Per questo motivo è stata scelta la balanced accuracy, che ovvia al problema della metrica precedente in quanto non è più indipendente dalla corretta classificazione di entrambe le classi in contemporanea.

Inoltre, per i migliori setting, sono stati calcolati altri parametri di scoring che possono aiutare a valutare la bontà del lavoro svolto; nello specifico:

- Recall, la metrica che indica la percentuale di positivi correttamente identificati dal modello analizzato, calcolata come in Equazione 4.

$$Recall = \frac{TP}{TP + FN}$$

Equazione 4. *Formula Recall*

- Precision, indica qual è la "confidence" con la quale il classificatore identifica un positivo (più è alta, più è probabile che un positivo sia effettivamente tale). La formula è riportata in Equazione 5.

$$Precision = \frac{TP}{TP + FP}$$

Equazione 5. *Formula Precision*

- Area Under Curve (AUC), misura l'area bidimensionale delimitata al di sotto della curva ROC. Essa rappresenta l'abilità di un classificatore di distinguere le classi, più è alta, migliori sono le performance del modello.
- Confusion Matrix, è una tabella che fornisce una rappresentazione dell'accuratezza di classificazione. Ogni colonna rappresenta i valori predetti, mentre ogni riga i valori reali; sulla diagonale principale sono riportate le corrette classificazioni.

Per procedere al tuning dei parametri, dunque, è stato sfruttato un oggetto della libreria "sklearn" di Python [25], nello specifico il "RandomizedSearchCV"/"GridSearchCV" a seconda del tipo di ricerca che si vuole svolgere. Esso prende in input il modello base del classificatore di cui si vogliono ottenere i migliori parametri, i dati su cui allenare tale classificatore con la relativa label, il dizionario contenente tutti i parametri da testare e il parametro di scoring prescelto per stabilire quale combinazione di parametri sia la migliore.

Per quanto riguarda i dati da fornire in input per questo processo di tuning, vista la scarsa numerosità di pazienti disponibili per questo studio si è deciso di inserire l'intero construction set; un ulteriore suddivisione in training e test set avrebbe ulteriormente ridotto i numeri a disposizione, compromettendo la solidità delle analisi condotte.

Avendo inserito l'intero construction set non suddiviso in training e test set, il tuning avviene internamente sfruttando una Stratified KFold. I dati facenti parte del training set sono splittati in training (70%) e test set (30%) per 5 ripetizioni (5 fold); ad ogni ripetizione cambiano i pazienti che compongono i due set e viene calcolato il valore del parametro di scoring scelto, per poi far la media su tutti i fold e ottenere il valor medio dello scoring.

In questo modo si è in grado di valutare la stabilità del classificatore allenato con quei parametri al variare dei training/test set a disposizione; se i risultati son stabili, si tradurranno in un valore medio dello scoring più alto, che di conseguenza porterà tale modello ad esser scelto come migliore.

Una volta esaurita la ricerca e individuato il modello migliore, si prosegue con una K-Fold Validation del modello "base" e del modello allenato con i parametri migliori, confrontando i parametri di scoring ottenuti; se il processo è andato a buon fine si noteranno dei valori più alti, sintomo del fatto che sono effettivamente stati trovati i parametri giusti per il setting attuale.

Una volta stabilito che il modello è il migliore possibile, si dovrebbe procedere con il training e la valutazione; per questo studio, si è deciso di effettuare uno step intermedio. Una volta che si allena un classificatore, esso può predire la classe di appartenenza dei dati fornitigli in input in due modalità:

- Predict, il classificatore fornirà in output la classe di appartenenza dei dati presi in input
- Predict_proba, in questo caso il classificatore fornirà una probabilità di appartenenza di tali dati alla classe 0 o alla classe 1. Tale probabilità, compresa tra 0 e 1, può tornare utile per fissare un threshold.

  Individuando un threshold, siamo in grado di ottimizzare ulteriormente la capacità predittiva del classificatore; sono stati testati vari threshold compresi tra 0 e 1 a passi di 0.1. Per ognuno di essi è stato calcolato il parametro di scoring prescelto, fino ad individuare il threshold che fornisse il valore più alto di tale parametro.

Tale tuning è stato svolto per ogni possibile combinazione di classificatore e feature selection, ottenendo così un totale di 25 "migliori" classificatori per il rischio (sei per tipologia di classificatore, uno per ogni metodo di FS) e 25 per la mutazione.

2.2.8 Validazione

I classificatori individuati al punto precedente son stati allenati sull'intero construction set per permettere al classificatore stesso di visionare più "esempi" possibili e fornire delle migliori prestazioni.

È risaputo che i classificatori dovrebbero essere allenati su dataset equilibrati al fine di non creare bias o overfitting; se questo fatto non risulta essere un problema per il classificatore del rischio, che presenta un construction set equilibrato, lo è invece nel caso del classificatore della mutazione. In questo caso, infatti, il construction risulta essere sbilanciato verso la classe 0

(mutazione KIT); un allenamento sull'intero construction potrebbe quindi essere vantaggioso in quanto il classificatore visiona più esempi da cui apprendere, ma potrebbe anche rivelarsi un fattore negativo dato che si potrebbe creare un bias verso la classe 0. Per questo motivo, si è deciso di effettuare entrambe le prove, vedendo quale scelta fornisse i risultati migliori.

Una volta allenati, tali classificatori sono stati testati sul training set e sul validation set esterno, ottenendo i parametri di scoring necessari per un confronto mirato ad individuare il miglior modello per la casistica presa in esame, ottenendo così i due classificatori definitivi.

3 Risultati

3.1 Classificazione del rischio di recidiva

Nelle seguenti tabelle sono riportati, per ogni tipologia di classificatore, i parametri di scoring ottenuti con i diversi metodi di Feature Selection sul training set e sul validation set per i classificatori del rischio di recidiva. Per selezionare il miglior accoppiamento Feature Selection/Classificatore, verranno osservati tutti i parametri, ponendo particolare attenzione sulle Confusion Matrix ottenute dalla validazione sul validation set esterno. Tali Confusion Matrix sono riportate con la classe vera sulle righe e la classe predetta sulle colonne.

Nella Tabella 6a è possibile osservare i parametri per il classificatore Bayesiano ottenuti sul construction set, nella 6b i parametri sul validation set. Non presenta nessun parametro da settare.

I Threshold migliori individuati sono: [0.1, 0.1, 0.1, 0.2, 0.1, 0.1]

Tabella 6a. *Risultati Bayesiano construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.7	0.41	0.9	0.68	0.75	0	23	1
							1	13	9
p-value	17	0.67	0.36	0.89	0.66	0.69	0	23	1
							1	14	8
ANOVA	37	0.7	0.41	0.9	0.68	0.76	0	23	1
							1	13	9
LASSO	4	0.72	0.45	0.91	0.71	0.82	0	23	1
							1	12	10
ANOVA + LASSO	4	0.7	0.5	0.79	0.69	0.82	0	21	3
							1	11	11
Importance	40	0.7	0.41	0.9	0.68	0.78	0	23	1
							1	13	9

Tabella 6b. *Risultati Bayesiano validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.67	0.65	0.81	0.68	0.76	0	7	3
							1	7	13
p-value	17	0.57	0.45	0.82	0.62	0.72	0	8	2
							1	11	9
ANOVA	37	0.67	0.6	0.86	0.7	0.76	0	8	2
							1	8	12
LASSO	4	0.67	0.65	0.81	0.68	0.74	0	7	3
							1	7	13
ANOVA + LASSO	4	0.57	0.5	0.77	0.6	0.67	0	7	3
							1	10	10
Importance	40	0.67	0.6	0.86	0.7	0.75	0	8	2
							1	8	12

Confrontando i vari parametri, è possibile notare come i risultati ottenuti sul construction set riportati in Tabella 6a siano molto simili per tutte le tipologie di feature selection analizzate; la precision risulta essere sempre molto alta (intorno al 90% in cinque casi su sei), ma accompagnata da una recall decisamente insoddisfacente.

Nel caso del Bayesiano, analizzando i parametri di scoring riportati in Tabella 6b, il miglior classificatore è ottenuto effettuando una feature selection ANOVA prendendo in considerazione le 37 features più utili.

In Tabella 7a/7b sono riportati i risultati relativi all'SVM. I thresholds sono: [0.5, 0.4, 0.5, 0.1, 0.5, 0.5].

Tabella 7a. *Risultati SVM construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.52	0	0	0.5	0.3	0	24	0
							1	22	0
p-value	17	0.7	0.64	0.7	0.69	0.72	0	18	6
							1	8	14
ANOVA	22	0.7	0.41	0.9	0.68	0.77	0	23	1
							1	13	9
LASSO	4	0.48	1	0.48	0.5	0.79	0	0	24
							1	0	22
ANOVA + LASSO	4	0.7	0.45	0.83	0.69	0.79	0	22	2
							1	12	10
Importance	64	0.74	0.45	1	0.73	0.75	0	24	0
							1	12	10

Tabella 7b. *Risultati SVM validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.33	0	0	0.5	0.34	0	10	0
							1	20	0
p-value	17	0.63	0.65	0.76	0.62	0.6	0	6	4
							1	7	13
ANOVA	22	0.73	0.7	0.88	0.75	0.74	0	8	2
							1	6	14
LASSO	4	0.67	1	0.67	0.5	0.76	0	0	10
							1	0	20
ANOVA + LASSO	4	0.57	0.5	0.77	0.6	0.6	0	7	3
							1	10	10
Importance	64	0.67	0.55	0.92	0.73	0.8	0	9	1
							1	9	11

Nel caso dell'SVM è possibile notare una variabilità più accentuata dei parametri ottenuti sul construction set; nella feature selection effettuata tramite LASSO e nel caso in cui non viene effettuata alcuna FS, infatti, il classificatore risulta assolutamente incapace di distinguere gli elementi appartenenti alle due classi. Tale comportamento è visibile anche successivamente nella validazione di tali classificatori sul validation set esterno.

In questo caso, il miglior metodo di Feature Selection risulta essere l'ANOVA con 22 features selezionate. I parametri migliori individuati durante il processo di tuning per questo modello sono:

- C: 100
- Decision_function_shape: ovo
- Degree: 2
- Gamma: auto
- Kernel: rbf
- Probability: True
- Shrinking: True

Nelle Tabelle 8a/8b sono stati inseriti i risultati relativi al kNN. Threshold: [0.4, 0.4, 0.4, 0.4, 0.6, 0.4]

Tabella 8a. *Risultati kNN construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	1	1	1	1	1	0	24	0
							1	0	22
p-value	17	0.83	0.73	0.89	0.82	0.9	0	22	2
							1	6	16
ANOVA	67	1	1	1	1	1	0	24	0
							1	0	22
LASSO	4	0.67	0.59	0.68	0.67	0.72	0	18	6
							1	9	13
ANOVA + LASSO	4	0.67	0.36	0.89	0.66	0.82	0	23	1
							1	14	8
Importance	62	1	1	1	1	1	0	24	0
							1	0	22

Tabella 8b. *Risultati kNN validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.57	0.4	0.89	0.65	0.63	0	9	1
							1	12	8
p-value	17	0.6	0.6	0.75	0.6	0.63	0	6	4
							1	8	12
ANOVA	67	0.8	0.85	0.85	0.77	0.89	0	7	3
							1	3	17
LASSO	4	0.7	0.85	0.74	0.62	0.72	0	4	6
							1	3	17
ANOVA + LASSO	4	0.6	0.45	0.9	0.68	0.64	0	9	1
							1	11	9
Importance	62	0.87	0.85	0.94	0.88	0.82	0	9	1
							1	3	17

Con questa tipologia di classificatore è possibile notare sul construction set un comportamento perfetto in tre casi su sei, portando il classificatore a classificare correttamente tutti gli esempi fornitigli in fase di allenamento.

Il miglior funzionamento è stato riscontrato sul metodo di FS basato sull'importance, che ha stavolta selezionato un totale di 62 features. Per quanto riguarda i parametri migliori individuati in fase di tuning, abbiamo:

- Weights: distance
- p: 1
- n_neighbors: 17
- metric: euclidean
- algorithm: auto

Nel caso del Random Forest non sono stati applicati metodi di feature selection, in quanto questa tipologia di modello seleziona automaticamente le migliori features ad ogni nodo. Per questo motivo, nelle Tabelle 9a/9b è riportata una sola riga di risultati relativi a questo classificatore: Threshold: 0.8

Tabella 9a. *Risultati Random Forest Construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	1	1	1	1	1	0	24	0
							1	0	22

Tabella 9b. *Risultati Random Forest validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.47	0.3	0.75	0.55	0.67	0	8	2
							1	14	6

Il Random Forest, in questo caso come in tutti i successivi, risulta essere un modello soggetto ad un'importante overfitting sui dati utilizzati per il suo allenamento. Questo classificatore, infatti, risulta essere sempre in grado di classificare in maniera perfetta i dati utilizzati per il suo allenamento ma, quando si tratta di dover classificare elementi a lui sconosciuti, non risulta essere in grado di generalizzare la sua capacità classificatoria.

I parametri migliori individuati sono:

- bootstrap: False
- criterion: entropy
- max_features: auto
- n_estimators: 75

Infine, nelle Tabelle 10a/10b sono riportati i risultati ottenuti con il classificatore XGBoost. Threshold: [0.7, 0.6, 0.2, 0.7, 0.1, 0.9]

Tabella 10a. *Risultati XGBoost construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.96	0.91	1	0.95	1	0	24	0
							1	2	20
p-value	17	0.98	0.95	1	0.98	1	0	24	0
							1	1	21
ANOVA	10	0.96	0.95	0.95	0.96	0.99	0	23	1
							1	1	21
LASSO	4	1	1	1	1	1	0	24	0
							1	0	22
ANOVA + LASSO	2	1	1	1	1	1	0	24	0
							1	0	22
Importance	45	1	1	1	1	1	0	24	0
							1	0	22

Tabella 10b. *Risultati XGBoost validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.53	0.45	0.75	0.57	0.67	0	7	3
							1	11	9
p-value	17	0.6	0.6	0.75	0.6	0.52	0	6	4
							1	8	12
ANOVA	10	0.6	0.65	0.72	0.57	0.48	0	5	5
							1	7	13
LASSO	4	0.53	0.5	0.71	0.55	0.61	0	6	4
							1	10	10
ANOVA + LASSO	2	0.57	0.75	0.65	0.48	0.54	0	2	8
							1	5	15
Importance	45	0.6	0.5	0.83	0.65	0.68	0	8	2
							1	10	10

Anche in questo caso, come già visto per il Random Forest, l'XGBoost tende in maniera importante all'overfitting performando in maniera quasi perfetta in tutti i metodi di FS sul construction ma riportando scarsi risultati in fase di validazione.

Il miglior metodo è risultato essere l'importance, con 45 features selezionate; nel tuning dei parametri, sono stati individuati i seguenti:

- subsample: 0.8

- n_estimators: 750
- min_samples_split: 6
- min_samples_leaf: 9
- max_features: log2
- max_depth: 6
- learning_rate: 0.05

Una volta individuati i migliori metodi di FS per ogni classificatore provato, è stato effettuato un confronto finale per individuare il migliore classificatore costruito per la valutazione del rischio. I dati di tali classificatori son stati riportati nelle Tabelle 11a/11b per permettere un confronto più agevole.

Tabella 11a. *Confronto migliori classificatori per tipologia construction*

Model + FS	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Bayesiano ANOVA	37	0.7	0.41	0.9	0.68	0.76	0	23	1
							1	13	9
SVM ANOVA	22	0.7	0.41	0.9	0.68	0.77	0	23	1
							1	13	9
kNN Importance	62	1	1	1	1	1	0	24	0
							1	0	22
Random Forest	86	1	1	1	1	1	0	24	0
							1	0	22
XGBoost Importance	45	1	1	1	1	1	0	24	0
							1	0	22

Tabella 11b. *Confronto migliori classificatori per tipologia validation*

Model + FS	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Bayesiano ANOVA	37	0.67	0.6	0.86	0.7	0.76	0	8	2
							1	8	12
SVM ANOVA	22	0.73	0.7	0.88	0.75	0.74	0	8	2
							1	6	14
kNN Importance	62	0.87	0.85	0.94	0.88	0.82	0	9	1
							1	3	17
Random Forest	86	0.47	0.3	0.75	0.55	0.67	0	8	2
							1	14	6
XGBoost Importance	45	0.6	0.5	0.83	0.65	0.68	0	8	2
							1	10	10

Osservando tali tabelle, il miglior classificatore risulta essere il kNN allenato sulle 62 features individuate tramite Importance; tale classificatore, infatti, presenta i parametri di scoring più alti di tutti gli altri sia sul construction set (a parità del Random Forest e dell'XGBoost, che ricordiamo soffrire di overfitting) che sul validation set, dove prevale la sua elevata capacità di generalizzazione ad elementi a lui sconosciuti.

Le 62 features selezionate da tale classificatore sono riportate in Tabella 12.

Tabella 12. *Features kNN Importance*

Gruppo	Features	Numero
Shape	Elongation, Flatness, LeastAxisLength, MahorAxisLength, Maximum2DDiameterColumn, Maximum2DDiameterRow, Maximum2DDiameterSlice, Maximum3DDiameter, MeshVolume, MinorAxisLength, Sphericity, SurfaceArea, SurfaceVolumeRatio, VoxelVolume	14
FirstOrder	10Percentile, Energy, Entropy, InterquartileRange, Kurtosis, Maximum, Mean, MeanAbsoluteDeviation, Median, Minimum, Range, RobustMeanAbsoluteDeviation, RootMeanSquared, TotalEnergy	14
GLCM	Autocorrelation, ClusterProminence, ClusterTendency, Correlation, DifferenceAverage, Id, Idm, Imc1, Imc2, JointAverage, JointEntropy, MCC, SumAverage, SumEntropy	14
GLDM	DependenceNonUniformity, DependenceNonUniformityNormalized, SmallDependenceEmphasis, SmallDependenceHighGrayLevelEmphasis	4
GLRLM	GrayLevelNonUniformityNormalized, GrayLevelVariance, HighGrayLevelRunEmphasis, RunEntropy, RunLengthNonUniformity, RunVariance, ShortRunHighGrayLevelEmphasis	7
GLSZM	GrayLevelNonUniformity, GrayLevelVariance, SizeZoneNonUniformity, SizeZoneNonUniformityNormalized, SmallAreaHighGrayLevelEmphasis, ZoneEntropy	6
NGTDM	Busyness, Coarseness, Strength	3

3.2 Classificazione mutazionale

In questo caso ci troviamo di fronte ad un construction set sbilanciato; si è deciso di testare due possibilità:

- Effettuare il tuning dei parametri sull'intero construction, per poi allenare il classificatore sull'intero construction set, anche se sbilanciato, in modo da fornire al classificatore più esempi possibili da cui apprendere
- Effettuare il tuning dei parametri sull'intero construction, per poi allenare il classificatore su un training set equilibrato ricavato dal construction; in questo modo il classificatore avrà meno esempi su cui allenarsi, ma si evita la possibile formazione di un bias verso la classe più numerosa

3.2.1 Allenamento sbilanciato

Come nel caso precedente, nelle seguenti tabelle sono stati riportati i risultati ottenuti con i diversi metodi di Feature Selection per ogni tipologia di classificatore analizzata.

Nelle Tabelle 13a/13b è possibile osservare tali parametri per il classificatore Bayesiano. Non presenta nessun parametro da settare. Threshold: [0.1, 0.2, 0.1, 0.9, 0.3, 0.9]

Tabella 13a. *Risultati Bayesiano construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.54	0.84	0.43	0.6	0.71	0	12	21
							1	3	16
p-value	2	0.71	0.68	0.59	0.71	0.68	0	24	9
							1	6	13
ANOVA	55	0.54	0.84	0.43	0.6	0.7	0	12	21
							1	3	16
LASSO	4	0.63	0	0	0.5	0.75	0	33	0
							1	19	0
ANOVA + LASSO	4	0.62	0.84	0.48	0.66	0.75	0	16	17
							1	3	16
Importance	6	0.71	0.42	0.67	0.65	0.66	0	29	4
							1	11	8

Tabella 13b. *Risultati Bayesiano validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.63	0.5	0.36	0.59	0.66	0	15	7
							1	4	4
p-value	2	0.73	0.62	0.5	0.7	0.79	0	17	5
							1	3	5
ANOVA	55	0.83	0.88	0.64	0.85	0.79	0	18	4
							1	1	7
LASSO	4	0.73	0	0	0.5	0.66	0	22	0
							1	8	0
ANOVA + LASSO	4	0.67	0.62	0.42	0.65	0.66	0	15	7
							1	3	5
Importance	6	0.87	0.62	0.83	0.79	0.69	0	21	1
							1	3	5

Rispetto al caso precedente del classificatore del rischio, il validation set risulta essere ancora più sbilanciato a favore della classe 0, rendendo di fatto molte metriche "pericolose" da analizzare senza rapportarle allo sbilanciamento del validation.

Nel caso del Bayesiano, il miglior classificatore è ottenuto effettuando una feature selection basata sul One Way ANOVA, che ha selezionato 55 features.

Nelle Tabelle 14a/14b sono riportati i risultati relativi all'SVM. Threshold: [0.1, 0.1, 0.3, 0.5, 0.4, 0.4].

Tabella 14a. *Risultati SVM construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.37	1	0.37	0.5	0.77	0	0	33
							1	0	19
p-value	2	0.37	1	0.37	0.5	0.77	0	0	33
							1	0	19
ANOVA	44	0.48	1	0.41	0.59	0.79	0	6	27
							1	0	19
LASSO	4	0.63	0	0	0.5	0.76	0	33	0
							1	19	0
ANOVA + LASSO	3	0.67	0.26	0.62	0.59	0.62	0	30	3
							1	14	5
Importance	5	0.79	0.58	0.79	0.74	0.75	0	30	3
							1	8	11

Tabella 14b. *Risultati SVM validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.27	1	0.27	0.5	0.56	0	0	22
							1	0	8
p-value	2	0.27	1	0.27	0.5	0.64	0	0	22
							1	0	8
ANOVA	44	0.8	0.88	0.58	0.82	0.88	0	17	5
							1	1	7
LASSO	4	0.77	0.12	1	0.56	0.74	0	22	0
							1	7	1
ANOVA + LASSO	3	0.8	0.62	0.62	0.74	0.71	0	19	3
							1	3	5
Importance	5	0.83	0.5	0.8	0.73	0.69	0	21	1
							1	4	4

Nel caso dell'SVM il miglior metodo di Feature Selection risulta essere nuovamente l'ANOVA, con 44 features selezionate. I parametri migliori individuati durante il processo di tuning per questo modello sono:

- C: 50
- Decision_function_shape: ovo
- Degree: 2
- Gamma: auto
- Kernel: sigmoid
- Probability: True
- Shrinking: True

Nelle Tabelle 15a/15b sono stati inseriti i risultati relativi al kNN. Threshold: [0.1, 0.4, 0.5, 0.4, /, 0.4].

Nel caso della feature selection combinata (ANOVA + LASSO) non è stato possibile procedere all'estrazione dei parametri di scoring, in quanto i due metodi di feature selection non hanno selezionato alcuna feature in comune.

Tabella 15a. *Risultati kNN construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.42	1	0.39	0.55	0.83	0	3	30
							1	0	19
p-value	2	0.85	0.74	0.82	0.82	0.89	0	30	3
							1	5	14
ANOVA	5	0.71	0.53	0.62	0.67	0.71	0	27	6
							1	9	10
LASSO	4	1	1	1	1	1	0	33	0
							1	0	19
ANOVA + LASSO	0	/	/	/	/	/	0	/	/
							1	/	/
Importance	3	0.73	0.47	0.69	0.68	0.69	0	29	4
							1	10	9

Tabella 15b. *Risultati kNN validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.3	1	0.28	0.52	0.57	0	1	21
							1	0	8
p-value	2	0.63	0.25	0.29	0.51	0.45	0	17	5
							1	6	2
ANOVA	5	0.87	0.62	0.83	0.79	0.67	0	21	1
							1	3	5
LASSO	4	0.4	0.75	0.27	0.51	0.71	0	6	16
							1	2	6
ANOVA + LASSO	0	/	/	/	/	/	0	/	/
							1	/	/
Importance	3	0.8	0.5	0.67	0.7	0.75	0	20	2
							1	4	4

Nel caso del kNN, il miglior funzionamento è stato riscontrato per la terza volta consecutiva sul metodo di FS basato su ANOVA, selezionando stavolta 5 features. Per quanto riguarda i parametri migliori individuati in fase di tuning, abbiamo:

- Weights: uniform
- p: 1
- n_neighbors: 7
- metric: chebyshev
- algorithm: auto

Nelle Tabelle 16a/16b sono riportati i risultati relativi al Random Forest. Threshold: [0.6].

Tabella 16a. *Risultati Random Forest construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	1	1	1	1	1	0	33	0
							1	0	19

Tabella 16b. *Risultati Random Forest validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.8	0.25	1	0.62	0.64	0	22	0
							1	6	2

Come già visto per il classificatore del rischio, il random forest tende a overfittare sul dataset utilizzato per il suo allenamento, riportando risultati perfetti nella classificazione del construction ma mostrando scarsa capacità di generalizzazione quando si tratta di dover classificare elementi mai visti prima. I parametri migliori individuati sono:

- bootstrap: False
- criterion: entropy
- max_features: sqrt
- n_estimators: 10

Infine, nelle Tabelle 17a/17b sono riportati i risultati ottenuti con il classificatore XGBoost. Threshold: [0.3, 0.6, 0.2, 0.2, 0.7, 0.2].

Tabella 17a. *Risultati XGBoost construction*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	1	1	1	1	1	0	33	0
							1	0	19
p-value	2	0.85	0.58	1	0.79	1	0	33	0
							1	8	11
ANOVA	10	0.65	1	0.51	0.73	0.93	0	15	18
							1	0	19
LASSO	4	1	1	1	1	1	0	33	0
							1	0	19
ANOVA + LASSO	1	0.85	0.58	1	0.79	0.95	0	33	0
							1	8	11
Importance	9	0.88	1	0.76	0.91	1	0	27	6
							1	0	19

Tabella 17b. *Risultati XGBoost validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.77	0.38	0.6	0.64	0.71	0	20	2
							1	5	3
p-value	2	0.77	0.12	1	0.56	0.62	0	22	0
							1	7	1
ANOVA	10	0.3	1	0.28	0.52	0.73	0	1	21
							1	0	8
LASSO	4	0.53	0.5	0.29	0.52	0.57	0	12	10
							1	4	4
ANOVA + LASSO	1	0.67	0.38	0.38	0.57	0.6	0	17	5
							1	5	3
Importance	9	0.77	0.62	0.56	0.72	0.68	0	18	4
							1	3	5

Anche per questo classificatore, come per il random forest, assistiamo ad un frequente overfitting. In questo caso le migliori performance sono ottenute con la feature selection per importance, selezionando 9 features. Nel tuning dei parametri, sono stati individuati i seguenti:

- subsample: 1
- n_estimators: 250
- min_samples_split: 6
- min_samples_leaf: 7
- max_features: log2

- max_depth: 3
- learning_rate: 0.01

Una volta individuati i migliori metodi di FS per ogni classificatore provato, è stato effettuato un confronto finale per individuare il migliore classificatore costruito per la classificazione della mutazione. I dati di tali classificatori son stati riportati nelle Tabelle 18a/18b per permettere un confronto più agevole.

Tabella 18a. *Confronto migliori classificatori per tipologia construction*

Model + FS	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Bayesiano ANOVA	55	0.54	0.84	0.43	0.6	0.7	0	12	21
							1	3	16
SVM ANOVA	44	0.48	1	0.41	0.59	0.79	0	6	27
							1	0	19
kNN ANOVA	5	0.71	0.53	0.62	0.67	0.71	0	27	6
							1	9	10
Random Forest	86	1	1	1	1	1	0	33	0
							1	0	19
XGBoost Importance	9	0.88	1	0.76	0.91	1	0	27	6
							1	0	19

Tabella 18b. *Confronto migliori classificatori per tipologia validation*

Model + FS	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Bayesiano ANOVA	55	0.83	0.88	0.64	0.85	0.79	0	18	4
							1	1	7
SVM ANOVA	44	0.8	0.88	0.58	0.82	0.88	0	17	5
							1	1	7
kNN ANOVA	5	0.87	0.62	0.83	0.79	0.67	0	21	1
							1	3	5
Random Forest	86	0.8	0.25	1	0.62	0.64	0	22	0
							1	6	2
XGBoost Importance	9	0.77	0.62	0.56	0.72	0.68	0	18	4
							1	3	5

Da tali tabelle è possibile notare come i due classificatori che soffrono di overfitting, il Random Forest e l'XGBoost, siano i più deludenti in fase di validazione, motivo per il quale vengono scartati.

Il classificatore che offre le migliori performances su entrambe le classi risulta essere il Bayesiano con feature selection ANOVA.

Le 55 features selezionate da tale classificatore sono riportate in Tabella 19.

Tabella 19. *Features Bayesiano ANOVA*

Gruppo	Features	Numero
Shape	MeshVolume, SurfaceArea, VoxelVolume	3
FirstOrder	InterquartileRange, MeanAbsoluteDeviation, Minimum, RobustMeanAbsoluteDeviation, TotalEnergy, Uniformity	6
GLCM	Autocorrelation, ClusterProminence, ClusterTendency, Correlation, DifferenceAverage, DifferenceEntropy, Id, Idm, Idmn, Idn, Imc1, Imc2, InverseVariance, JointAverage, MCC, SumAverage, SumEntropy	17
GLDM	DependenceEntropy, DependenceNonUniformity, DependenceNonUniformityNormalized, DependenceVariance, HighGrayLevelEmphasis, LargeDependenceEmphasis, SmallDependenceEmphasis	7
GLRLM	GrayLevelNonUniformityNormalized, HighGrayLevelRunEmphasis, LongRunEmphasis, LongRunHighGrayLevelEmphasis, RunEntropy, RunLengthNonUniformity, RunLengthNonUniformityNormalized, RunPercentage, RunVariance, ShortRunEmphasis	10
GLSZM	GrayLevelNonUniformity, GrayLevelNonUniformityNormalized, GrayLevelVariance, HighGrayLevelZoneEmphasis, SizeZoneNonUniformity, SizeZoneNonUniformityNormalized, SmallAreaEmphasis, ZoneEntropy, ZonePercentage	9
NGTDM	Busyness, Complexity, Strength	3

3.2.2 Allenamento bilanciato

Ripetiamo lo stesso procedimento, utilizzando per l'allenamento dei classificatori un training equilibrato ricavato dal construction set.

Nelle Tabelle 20a/20b è possibile osservare i risultati ottenuti per il classificatore Bayesiano. Non presenta nessun parametro da settare. Threshold: [0.1, 0.4, 0.1, 0.5, 0.1, 0.9].

Tabella 20a. *Risultati Bayesiano training equilibrato*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.58	0.79	0.56	0.58	0.76	0	7	12
							1	4	15
p-value	2	0.68	0.53	0.77	0.68	0.78	0	16	3
							1	9	10
ANOVA	33	0.61	0.74	0.58	0.61	0.75	0	9	10
							1	5	14
LASSO	4	0.71	0.74	0.7	0.71	0.8	0	13	6
							1	5	14
ANOVA + LASSO	3	0.58	1	0.54	0.58	0.76	0	3	16
							1	0	19
Importance	12	0.68	0.63	0.71	0.68	0.74	0	14	5
							1	7	12

Tabella 20b. *Risultati Bayesiano validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.73	0.75	0.5	0.74	0.74	0	16	6
							1	2	6
p-value	2	0.73	0.5	0.5	0.66	0.68	0	18	4
							1	4	4
ANOVA	33	0.8	0.75	0.6	0.78	0.77	0	18	4
							1	2	6
LASSO	4	0.6	0.62	0.36	0.61	0.67	0	13	9
							1	3	5
ANOVA + LASSO	3	0.27	1	0.27	0.5	0.8	0	0	22
							1	0	8
Importance	12	0.87	0.75	0.75	0.83	0.76	0	20	2
							1	2	6

In questo caso, il miglior classificatore è ottenuto effettuando una feature selection basata sull'Importance, selezionando 12 features.

Nelle Tabelle 21a/21b sono riportati i risultati relativi all'SVM. Threshold: [0.4, 0.4, 0.4, 0.1, 0.4, 0.4].

Tabella 21a. *Risultati SVM training equilibrato*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.68	1	0.61	0.68	1	0	7	12
							1	0	19
p-value	2	0.74	0.74	0.74	0.74	0.81	0	14	5
							1	5	14
ANOVA	38	0.71	0.84	0.67	0.71	0.81	0	11	8
							1	3	16
LASSO	4	0.5	1	0.5	0.5	0.8	0	0	19
							1	0	19
ANOVA + LASSO	2	0.58	0.95	0.55	0.58	0.77	0	4	15
							1	1	18
Importance	36	1	1	1	1	1	0	19	0
							1	0	19

Tabella 21b. *Risultati SVM validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.53	0.38	0.25	0.48	0.56	0	13	9
							1	5	3
p-value	2	0.5	0.88	0.33	0.62	0.69	0	8	14
							1	1	7
ANOVA	38	0.73	0.75	0.5	0.74	0.76	0	16	6
							1	2	6
LASSO	4	0.27	1	0.27	0.5	0.72	0	0	22
							1	0	8
ANOVA + LASSO	2	0.43	0.88	0.3	0.57	0.83	0	6	16
							1	1	7
Importance	36	0.6	1	0.4	0.73	0.73	0	10	12
							1	0	8

Nel caso dell'SVM il miglior metodo di Feature Selection risulta essere l'ANOVA con 38 features selezionate. I parametri migliori individuati durante il processo di tuning per questo modello sono:

- C: 100
- Decision_function_shape: ovo
- Degree: 2
- Gamma: auto
- Kernel: sigmoid

- Probability: True
- Shrinking: True

Nelle Tabelle 22a/22b sono stati inseriti i risultati relativi al kNN. Threshold: [0.2, 0.4, 0.4, 0.5, 0.6, 0.5].

Tabella 22a. *Risultati kNN training equilibrato*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.68	0.95	0.62	0.68	0.84	0	8	11
							1	1	18
p-value	2	0.84	0.79	0.88	0.84	0.91	0	17	2
							1	4	15
ANOVA	18	1	1	1	1	1	0	19	0
							1	0	19
LASSO	4	1	1	1	1	1	0	19	0
							1	0	19
ANOVA + LASSO	1	1	1	1	1	1	0	19	0
							1	0	19
Importance	1	0.68	0.47	0.82	0.68	0.78	0	17	2
							1	10	9

Tabella 22b. *Risultati kNN validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.33	0.88	0.27	0.51	0.61	0	3	19
							1	1	7
p-value	2	0.33	0.5	0.2	0.39	0.41	0	6	16
							1	4	4
ANOVA	18	0.77	0.88	0.54	0.8	0.8	0	16	6
							1	1	7
LASSO	4	0.67	0.62	0.42	0.65	0.76	0	15	7
							1	3	5
ANOVA + LASSO	1	0.73	0.5	0.5	0.66	0.76	0	18	4
							1	4	4
Importance	1	0.77	0.62	0.56	0.72	0.75	0	18	4
							1	3	5

Stavolta il kNN ha mostrato qualche caso di overfitting nel corso dell'addestramento, non riuscendo a performare in maniera altrettanto adeguata nel validation set. Il miglior risultato è

stato riscontrato utilizzando per la feature selection l'ANOVA, che ha selezionato 18 features. Per quanto riguarda i parametri migliori individuati in fase di tuning, abbiamo:

- Weights: distance
- p: 1
- n_neighbors: 17
- metric: manhattan
- algorithm: auto

Nelle Tabelle 23a/23b sono riportati i risultati relativi al Random Forest. Threshold: [0.4].

Tabella 23a. *Risultati Random Forest training equilibrato*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	1	1	1	1	1	0	19	0
							1	0	19

Tabella 23b. *Risultati Random Forest validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.73	0.62	0.5	0.7	0.74	0	17	5
							1	3	5

I parametri migliori individuati sono:

- bootstrap: True
- criterion: entropy
- max_features: log2
- n_estimators: 10

Infine, nelle Tabelle 24a/24b sono riportati i risultati ottenuti con il classificatore XGBoost. Threshold: [0.5, 0.3, 0.2, 0.3, 0.3, 0.1].

Tabella 24a. *Risultati XGBoost training equilibrato*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	1	1	1	1	1	0	19	0
							1	0	19
p-value	2	0.82	1	0.73	0.82	0.95	0	12	7
							1	0	19
ANOVA	10	0.58	1	0.54	0.58	0.92	0	3	16
							1	0	19
LASSO	4	1	1	1	1	1	0	19	0
							1	0	19
ANOVA + LASSO	1	0.84	0.89	0.81	0.84	0.93	0	15	4
							1	2	17
Importance	7	1	1	1	1	1	0	19	0
							1	0	19

Tabella 24b. *Risultati XGBoost validation*

Feature Selection	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
/	86	0.7	0.25	0.4	0.56	0.74	0	19	3
							1	6	2
p-value	2	0.53	0.75	0.33	0.6	0.7	0	10	12
							1	2	6
ANOVA	10	0.27	1	0.27	0.5	0.78	0	0	22
							1	0	8
LASSO	4	0.63	0.75	0.4	0.67	0.75	0	13	9
							1	2	6
ANOVA + LASSO	1	0.53	0.75	0.33	0.6	0.58	0	10	12
							1	2	6
Importance	7	0.67	0.62	0.42	0.65	0.62	0	15	7
							1	3	5

In questo caso le migliori performance sono ottenute con la feature selection eseguita con il LASSO, che ha selezionato 4 features. Nel tuning dei parametri, sono stati individuati i seguenti:

- subsample: 1
- n_estimators: 250
- min_samples_split: 2
- min_samples_leaf: 9
- max_features: log2

- max_depth: 2
- learning_rate: 0.1

Una volta individuati i migliori metodi di FS per ogni classificatore provato, è stato effettuato un confronto finale per individuare il migliore classificatore costruito per la classificazione della mutazione. I dati di tali classificatori son stati riportati nelle Tabelle 25a/25b per permettere un confronto più agevole.

Tabella 25a. *Confronto migliori classificatori per tipologia training equilibrato*

Model + FS	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Bayesiano Importance	12	0.68	0.63	0.71	0.68	0.74	0	14	5
							1	7	12
SVM ANOVA	38	0.71	0.84	0.67	0.71	0.81	0	11	8
							1	3	16
kNN ANOVA	18	1	1	1	1	1	0	19	0
							1	0	19
Random Forest	86	1	1	1	1	1	0	19	0
							1	0	19
XGB LASSO	4	1	1	1	1	1	0	19	0
							1	0	19

Tabella 25b. *Confronto migliori classificatori per tipologia validation*

Model + FS	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Bayesiano Importance	12	0.87	0.75	0.75	0.83	0.76	0	20	2
							1	2	6
SVM ANOVA	38	0.73	0.75	0.5	0.74	0.76	0	16	6
							1	2	6
kNN ANOVA	18	0.77	0.88	0.54	0.8	0.8	0	16	6
							1	1	7
Random Forest	86	0.73	0.62	0.5	0.7	0.74	0	17	5
							1	3	5
XGB LASSO	4	0.63	0.75	0.4	0.67	0.75	0	13	9
							1	2	6

Osservando tali tabelle, è possibile notare ancora una volta la tendenza del random forest e dell'XGB all'overfitting, comportamento che stavolta si è esteso anche al kNN, che riesce comunque a mantenere ottimi risultati anche in fase di validazione dimostrando di aver effettuato un allenamento soddisfacente.

In questo caso due classificatori hanno ottenuto performance confrontabili, il Bayesiano con metodo di FS per importance e il kNN con ANOVA. È possibile notare come la recall sia notevolmente a favore del kNN, ma questo è dovuto al numero bassissimo di pazienti appartenenti alla classe 1. In compenso, il bayesiano fornisce una precision molto più alta, essendo in grado di classificare in maniera migliore i pazienti appartenenti alla classe 0.

Tenendo conto dello scopo per cui è stato creato questo classificatore, si è deciso di scegliere il kNN, in quanto più performante nella classificazione dei pazienti appartenenti alla classe delle mutazioni non responsive all'imatinib; in questo modo si può evitare la somministrazione di un farmaco che non avrebbe alcun effetto. Il minor valore di precision dovrà esser compensato da un'analisi più approfondita che permetta di stabilire in maniera definitiva l'appartenenza alla classe corretta.

Le 18 features selezionate dal kNN con metodo di FS ANOVA sono riportate in Tabella 26.

Tabella 26. *Features kNN ANOVA*

Gruppo	Features	Numero
Shape	/	0
FirstOrder	/	0
GLCM	Autocorrelation, ClusterProminence, ClusterTendency, Correlation, DifferenceAverage, DifferenceEntropy, Idmn, Idn, Imc1, Imc2, InverseVariance, MCC, SumEntropy	13
GLDM	DependenceEntropy	1
GLRLM	LongRunHighGrayLevelEmphasis, RunEntropy	2
GLSZM	ZoneEntropy	1
NGTDM	Complexity	1

4 Discussione

In questo studio è stata valutata la possibilità di costruire due classificatori che, allenati su features estratte da immagini TC di pazienti affetti da GISTs, siano in grado di predire la stratificazione del rischio di recidiva secondo Miettinen/AFIP e il profilo mutazionale di questa tipologia di tumori.

Per quanto riguarda il classificatore del rischio, dopo varie prove è stato individuato come miglior modello un kNN allenato su 62 features selezionate tramite Importance, i cui risultati su construction e validation set sono riportati in Tabella 27.

Tabella 27. *Risultati kNN classificazione del rischio*

Set	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Construction	62	1	1	1	1	1	0	24	0
							1	0	22
Validation	62	0.87	0.85	0.94	0.88	0.82	0	9	1
							1	3	17

Analizzando nello specifico questi risultati, possiamo dire che:

- Construction Set. Il classificatore allenato sull'intero construction riesce successivamente a riconoscere alla perfezione tutti gli elementi fornitigli in fase di training. Questo significa che il modello selezionato ha imparato in maniera efficace dagli esempi presi in input ed è in grado di individuare le caratteristiche necessarie per distinguere gli elementi appartenenti alle due classi differenti; una classificazione così accurata potrebbe far pensare ad un overfitting sui dati, ma come si vedrà nel validation set, il classificatore presenta un'ottima capacità di generalizzazione, riuscendo ad ottenere buoni risultati anche su elementi non utilizzati in fase di addestramento.
- Accuracy/Balanced accuracy 87%. L'accuracy è la percentuale di totali corretti classificati. Essendo uno studio per la stratificazione del rischio di recidiva tumorale, parametro sul quale verrà poi stabilita la strategia terapeutica da adottare, è fondamentale avere un classificatore che sia in grado di mantenere un'elevata capacità di corretti classificati, in quanto un errore potrebbe significare la morte dei pazienti. L'87% è un buon valore, ma vista l'importanza dell'applicazione è ritenuto tuttora non sufficiente per un utilizzo effettivo nella vita reale.

- Recall 85%. Il recall è la percentuale di pazienti appartenenti alla classe 1 correttamente classificati. Anche in questo caso ci si trova di fronte ad un buon valore, ma non sufficiente per questa tipologia di studio, in quanto significherebbe contemplare la possibilità di classificare il 15% dei pazienti ad alto rischio come pazienti a basso rischio, compromettendo la terapia successiva e determinando esiti negativi.
- Precision 94%. La precision è la percentuale con la quale un paziente classificato come appartenente alla classe 1 vi appartenga per davvero. Con questo classificatore vengono classificati 18 pazienti appartenenti alla classe 1; di questi 18, ben 17 ne fanno effettivamente parte. Quando questo classificatore riconosce un paziente come appartenente alla classe 1, e di conseguenza ad alto rischio, si può ritenere tale predizione accurata.
- AUC 82%. L'Area Under the Curve è un parametro di accuratezza. Per un valore del 50% il test non è considerabile informativo, per valori via via crescenti dal 50 al 100% il test assume sempre più significatività.

Nel caso di classificazione del profilo mutazione sono stati individuati due migliori classificatori: uno allenato con l'intero construction set (fortemente sbilanciato verso la classe 0) e uno allenato su un sottoinsieme del construction avente una numerosità delle classi equilibrata.

In Tabella 28 e 29 sono riportati rispettivamente i risultati ottenuti nel classificatore sbilanciato e bilanciato; il primo è un bayesiano allenato su 55 features individuate con l'ANOVA, il secondo un kNN allenato con 18 features sempre selezionate tramite ANOVA.

Tabella 28. *Risultati bayesiano classificazione profilo mutazionale sbilanciata*

Set	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Construction	55	0.54	0.84	0.43	0.6	0.7	0	12	21
							1	3	16
Validation	55	0.83	0.88	0.64	0.85	0.79	0	18	4
							1	1	7

Tabella 29. *Risultati kNN classificazione profilo mutazionale bilanciata*

Set	# features	Accuracy	Recall	Precision	Balanced Accuracy	AUC	Confusion Matrix		
								0	1
Construction	18	1	1	1	1	1	0	19	0
							1	0	19
Validation	18	0.77	0.88	0.54	0.8	0.8	0	16	6
							1	1	7

In questo caso è possibile effettuare le seguenti osservazioni:

- Construction Set sbilanciato/Training Set equilibrato. Confrontando i risultati ottenuti sul construction set in entrambi i classificatori, è facile osservare come nel secondo caso (classificatore allenato su training set equilibrato) i risultati siano perfetti, mentre nel primo caso risultino essere piuttosto deludenti. Questo può voler significare che, avendo pochi esempi a disposizione e per lo più sbilanciato verso una classe, il classificatore non sia stato in grado di cogliere in maniera efficace le caratteristiche distintive tra le due classi, rendendo di fatto il modello incapace di distinguerle. Come si vedrà poi in seguito, però, tale comportamento non si riflette in fase di validazione, dove il modello sarà in grado di generalizzare adeguatamente in entrambi i casi
- Accuracy 83%/77%. In una situazione come questa, in cui il validation set risulta essere sbilanciato a favore della classe 0 (mutazione KIT), tale metrica risulta essere poco significativa. Una corretta classificazione della classe più numerosa, infatti, potrebbe "oscurare" un cattivo funzionamento su quella meno numerosa mantenendo un valore elevato. Saranno fatte considerazioni più significative sulla balanced accuracy, che in questo caso risulta essere più adatta allo scopo
- Recall 88%/88%. In entrambi i classificatori, la recall assume la stessa percentuale. Essi sono infatti in grado di riconoscere 7 pazienti su 8 effettivamente appartenenti alla classe 1; ricordando lo scopo del classificatore, è importante individuare in maniera corretta tali pazienti per evitare la somministrazione di un farmaco che non avrebbe effetti positivi per il decorso della malattia.
- Precision 64%/54%. Leggero miglioramento nel caso del classificatore allenato sull'intero construction set, ma percentuali molto basse in entrambi i casi; in un utilizzo reale, se un paziente fosse classificato come appartenente alla classe 1, sarebbero comunque necessarie ulteriori verifiche per accertarne l'effettiva appartenenza, dato che entrambi i classificatori risultano poco precisi.

- Balanced Accuracy 85%/80%. Valore leggermente favorevole al primo classificatore, che presenta una miglior classificazione per la classe 0 rispetto al secondo.
- AUC 79%/80%. Valori assolutamente paragonabili tra di loro

Dopo aver confrontato i due classificatori del profilo mutazionale, è stato scelto come migliore il bayesiano allenato sull'intero construction set visto il leggero vantaggio sul kNN in termini di Precision e Balanced Accuracy. Nonostante lo sbilanciamento del dataset utilizzato in fase di training, tale svantaggio è stato compensato dal maggior numero di esempi portati in dote al classificatore; vista l'esigua numerosità dei pazienti arruolati per questo studio, infatti, era fondamentale riuscire ad allenare il modello su quanti più elementi possibile per permettergli di osservare più caratteristiche e performare meglio in fase di validazione.

Volendo confrontare i risultati ottenuti con i progressi attuali della comunità scientifica, è stata svolta una ricerca atta ad individuare studi simili a questo, aventi come obiettivo la stratificazione del rischio di recidiva o l'individuazione del profilo mutazionale tramite machine learning. Tale ricerca ha evidenziato una notevole scarsità di lavori scientifici inerenti all'argomento, che risulta essere al momento un territorio ancora inesplorato; dal punto di vista mutazionale non sono stati trovati studi che includessero il machine learning. Per quanto riguarda la stratificazione del rischio, invece, sono stati individuati alcuni studi, di cui solamente due multicentrici come quello condotto in questo lavoro di tesi.

Nel lavoro svolto da Minhong Wang et al. [26] si voleva stabilire il rischio di recidiva secondo l'NIH modified Criteria basandosi su features radiomiche; in tale studio sono stati arruolati 324 pazienti provenienti da tre centri differenti, costruendo un training di 180 pazienti dal centro più grosso e un validation set di 144 pazienti affluenti ai restanti due centri. In questo lavoro è stata utilizzata come feature selection l'importance, selezionando solamente le prime 10 features da provare su tre tipologie diverse di classificatori: logistic regression, svm e random forest. I risultati ottenuti sono stati valutati tramite AUC, il random forest è risultato essere il classificatore migliore riportando per il training set 88% AUC e per il validation set 90% AUC. Confrontando tali valori con i risultati ottenuti in questo studio, è possibile osservare come nel training set il classificatore individuato risulta più performante di quello dello studio (100% vs 88%), mentre nel validation set risulti essere un po' meno capace (82% vs 90%). Nel lavoro di tesi qui presentato, inoltre, son stati analizzati molti più metodi di FS e tipologie di classificatori, riportando più metriche di scoring e svolgendo un'analisi più completa.

Nel secondo studio individuato nella ricerca bibliografica svolto da Yancheng Song et al. [27] si è sempre posto come obiettivo la stratificazione del rischio sfruttando le features radiomiche; in tale lavoro sono stati arruolati 500 pazienti provenienti da due centri, formando un training set di 346 pazienti e un validation set di 154. In questo caso sono state analizzate due tipologie di feature selection (mRMR e LASSO) prendendo le prime 20 features e testandole su 5 diversi classificatori: logistic regression, decision tree, random forest, svm ed adaboost. Il miglior risultato si è individuato con l'SVM, riportando un 89.5% AUC sul training e un 84.7% sul validation, con relativa accuracy dell'81.8% sul training e dell'81.2% sul validation. I risultati individuati dal lavoro di tesi precedentemente presentati risultano essere assolutamente confrontabili con quelli individuati dal presente lavoro in termini di AUC e addirittura superiori in accuracy.

In conclusione, i risultati ottenuti in questo lavoro di tesi non sono sufficienti per sostituire un prelievo istologico in pazienti affetti da GIST, ma sono un primo passo verso una ricerca che potrebbe rendere le features radiomiche dei nuovi parametri prognostici grazie ai quali ottenere delle indicazioni sull'aggressività della malattia, sulla mutazione genetica e, di conseguenza, sulla linea terapeutica adattata al singolo paziente.

Un fattore che ha influenzato notevolmente i risultati di questo studio è stata la scarsità campionaria messa a disposizione; trattandosi di un tumore raro, infatti, risulta molto difficile raccogliere dati da sottoporre ad analisi. I classificatori, al fine di raggiungere delle buone performances, necessitano di quanti più campioni possibili da cui "imparare"; è facilmente comprensibile che, se allenati su poche decine di campioni, non siano in grado di ottenere una conoscenza sufficiente ad ottenere buoni risultati su campioni mai visti prima.

Una conseguenza della scarsità campionaria risulta essere l'instabilità dei parametri di scoring ricavati dal presente lavoro di tesi; avendo in alcuni casi pochissimi pazienti appartenenti ad una classe, la corretta/errata classificazione di uno solo di loro poteva comportare una differenza nel parametro anche del 15/20%, rendendo di fatto molte metriche difficilmente interpretabili.

Un ulteriore fattore negativo che ha influenzato il classificatore del profilo mutazionale è stato il forte sbilanciamento dei dataset a favore della mutazione KIT, come era facilmente prevedibile, sapendo che tale mutazione risulta essere la più presente tra i GISTs (>80%). Per allenare un classificatore si tende sempre ad utilizzare dei dataset equilibrati; in questo caso si

è dovuto allenare il classificatore del profilo mutazionale con un construction set sbilanciato, portando irrimediabilmente ad un peggioramento delle sue performances.

Effettuare studi analoghi con una numerosità campionaria maggiore è necessario per indagare la possibilità di sfruttare le features radiomiche all'interno di modelli di supporto decisionali, eliminando la necessità di dover svolgere un prelievo istologico nei pazienti affetti da GIST.

Bibliografia

[1] CASALI, P. G., et al. Gastrointestinal stromal tumours: ESMO–EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology, 2018, 29: iv68-iv78.

[2] MACHAIRIOTIS, Nikolaos, et al. Gastrointestinal stromal tumor mesenchymal neoplasms: the offspring that choose the wrong path. Journal of multidisciplinary healthcare, 2013, 6: 127.

[3] CORLESS, Christopher L.; BARNETT, Christine M.; HEINRICH, Michael C. Gastrointestinal stromal tumours: origin and molecular oncology. Nature Reviews Cancer, 2011, 11.12: 865-878.

[4] MIETTINEN, Markku; LASOTA, Jerzy. Gastrointestinal stromal tumors. Gastroenterology Clinics, 2013, 42.2: 399-415.

[5] AIOM. Linee guida SARCOMI DEI TESSUTI MOLLI E GIST.

[6] ZHANG, Haixin; LIU, Qi. Prognostic indicators for gastrointestinal stromal tumors: a review. Translational Oncology, 2020, 13.10: 100812.

[7] MIETTINEN, Markku; LASOTA, Jerzy. Gastrointestinal stromal tumors: review on morphology, molecular pathology, prognosis, and differential diagnosis. Archives of pathology & laboratory medicine, 2006, 130.10: 1466-1478.

[8] JOENSUU, Heikki. Gastrointestinal stromal tumor (GIST). Annals of Oncology, 2006, 17: x280-x286.

[9] MANTESE, George. Gastrointestinal stromal tumor: epidemiology, diagnosis, and treatment. Current opinion in gastroenterology, 2019, 35.6: 555-559.

[10] MIETTINEN, Markku; LASOTA, Jerzy. Gastrointestinal stromal tumors: pathology and prognosis at different sites. In: Seminars in diagnostic pathology. WB Saunders, 2006. p. 70-83.

[11] NILSSON, Bengt, et al. Gastrointestinal stromal tumors: the incidence, prevalence, clinical course, and prognostication in the preimatinib mesylate era: a population-based study in western Sweden. Cancer, 2005, 103.4: 821-829.

[12] KIKUCHI, Hirotoshi; KONNO, Hiroyuki; TAKEUCHI, Hiroya. Risk Classification. In: Gastrointestinal Stromal Tumor. Springer, Singapore, 2019. p. 61-77.

[13] TETSUHITO, Muranaka; KOMATSU, Yoshito. Treatment Guidelines. In: Gastrointestinal Stromal Tumor. Springer, Singapore, 2019. p. 79-87.

[14] SEPE, Paul S.; BRUGGE, William R. A guide for the diagnosis and management of gastrointestinal stromal cell tumors. Nature reviews Gastroenterology & hepatology, 2009, 6.6: 363-371.

[15] SANDRASEGARAN, Kumaresan, et al. Gastrointestinal stromal tumors: CT and MRI findings. European radiology, 2005, 15.7: 1407-1414.

[16] GHANEM, Nadir, et al. Computed tomography in gastrointestinal stromal tumors. European radiology, 2003, 13.7: 1669-1678.

[17] HORTON, Karen M., et al. Computed tomography imaging of gastrointestinal stromal tumors with pathology correlation. Journal of computer assisted tomography, 2004, 28.6: 811-817.

[18] TAKAO, Hidemasa, et al. Gastrointestinal stromal tumor of the retroperitoneum: CT and MR findings. European radiology, 2004, 14.10: 1926-1929.

[19] VAN GRIETHUYSEN, Joost JM, et al. Computational radiomics system to decode the radiographic phenotype. Cancer research, 2017, 77.21: e104-e107.

[20] BERRAR, Daniel. Bayes' theorem and naive Bayes classifier. Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics; Elsevier Science Publisher: Amsterdam, The Netherlands, 2018, 403-412.

[21] PISNER, Derek A.; SCHNYER, David M. Support vector machine. In: Machine Learning. Academic Press, 2020. p. 101-121.

[22] PETERSON, Leif E. K-nearest neighbor. Scholarpedia, 2009, 4.2: 1883.

[23] CUTLER, Adele; CUTLER, D. Richard; STEVENS, John R. Random forests. In: Ensemble machine learning. Springer, Boston, MA, 2012. p. 157-175.

[24] CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

[25] PEDREGOSA, Fabian, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011, 12: 2825-2830.

[26] WANG, Minhong, et al. Computed-Tomography-Based Radiomics Model for Predicting the Malignant Potential of Gastrointestinal Stromal Tumors Preoperatively: A Multi-Classifier and Multicenter Study. Frontiers in Oncology, 2021, 11.

[27] SONG, Yancheng, et al. Radiomics Nomogram Based on Contrast-enhanced CT to Predict the Malignant Potential of Gastrointestinal Stromal Tumor: A Two-center Study. Academic Radiology, 2021.