UNIVERSITÀ
DI PAVIA

FACOLTA' DI INGEGNERIA
DIPARTIMENTO DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN BIOINGEGNERIA

TESI DI LAUREA

TITOLO:

TABULAR FOUNDATION MODELS PER LA PREDIZIONE DELLA DIAGNOSI FIBROTICA NELLE NEOPLASIE MIELOPROLIFERATIVE CON UN APPROCCIO BASATO SULL'INTEGRAZIONE DI DATI CLINICI E GENOMICI

Candidato: Ylenia Anna Pia Esposto

Relatore: Prof.ssa Arianna Dagliati
Correlatori: Prof.ssa Giovanna Nicora, Simone Rancati

A.A. 2025/2026

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Introduzione

In questo capitolo si introdurrà il contesto clinico e tecnologico alla base del presente lavoro di tesi. Verranno presentate le neoplasie mieloproliferative (MPN) e il problema legato alla diagnosi del fenotipo fibrotico.

Verranno inoltre illustrati gli obiettivi specifici della tesi, incentrati sullo sviluppo e valutazione di modelli predittivi in grado di identificare i pazienti con fibrosi.

1.1 Contesto

Le neoplasie mieloproliferative (MPN) sono un gruppo eterogeneo di malattie clonali croniche della cellula staminale emopoietica, caratterizzate dalla proliferazione persistente di una o più linee cellulari mieloidi, che porta a un conseguente aumento del numero di eritrociti (globuli rossi), leucociti (globuli bianchi) e trombociti (piastrine).

Esse si distinguono in base alla presenza o all'assenza del riarrangiamento cromosomico Philadelphia. L'unica MPN Philadelphia positive (Ph+) è la leucemia mieloide cronica (LMC), caratterizzata dalla traslocazione reciproca tra i cromosomi 9 e 22, t(9;22). Questa traslocazione determina la formazione del gene di fusione BCR::ABL1 e genera un cromosoma 22 alterato, denominato cromosoma Philadelphia (Ph) [1].

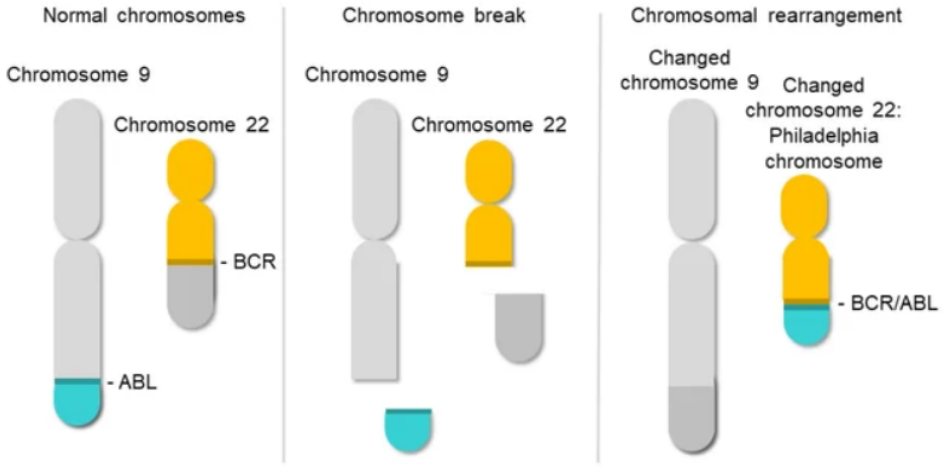

Figura 1.1: Struttura dell'oncogene BCR/ABL1. Rappresentazione schematica della traslocazione t(9;22)(q34;q11) che dà origine al cromosoma Philadelphia.

Le MPN Ph+ sono patologie che non presentano particolari difficoltà nella diagnosi proprio grazie al gene BCR-ABL1: è sufficiente, infatti, evidenziarne la presenza con un semplice esame FISH (Fluorescence In Situ Hybridization, un test genetico che permette di individuare specifiche sequenze di DNA sui cromosomi attraverso l'uso di sonde fluorescenti).

Le altre principali MPN classiche, come policitemia vera (PV), trombocitemia essenziale (TE) e mielofibrosi (MF), sono invece Philadelphia negative (Ph−). La loro distinzione risulta difficoltosa: queste patologie, infatti, condividono alcune caratteristiche cliniche e molti parametri ematici possono essere secondari a condizioni reattive (es. carenze di ferro, infiammazioni).

Particolare rilevanza assume, in questo contesto, l'identificazione del fenotipo fibrotico. La progressiva fibrosi del midollo osseo [2] costituisce, infatti, una caratteristica fondamentale della mielofibrosi. L'unico esame in grado di dirimere ogni dubbio è la biopsia osteomidollare (BOM), ovvero la raccolta di un campione di midollo osseo, che però rappresenta una metodica invasiva e dispendiosa, sia in termini economici che temporali.

1.2 IA in medicina

Le caratteristiche delle MPN Ph− e le difficoltà legate alla diagnosi e alla caratterizzazione del fenotipo fibrotico, rendono queste patologie un contesto particolarmente interessante per l'applicazione di metodologie di intelligenza artificiale (IA). Negli ultimi anni, infatti, l'IA ha assunto un ruolo crescente in medicina, grazie alla disponibilità sempre maggiore di dati clinici, laboratoristici, genetici e di imaging. I modelli di apprendimento automatico possono essere utilizzati per individuare relazioni complesse tra variabili, supportare la diagnosi, stimare la prognosi e identificare sottogruppi di pazienti caratterizzati da differenti profili di rischio.

Nelle prime applicazioni del machine learning in ambito clinico, l'approccio più comune consisteva nell'addestramento di modelli supervisionati specifici per un singolo compito, utilizzando un insieme predefinito di variabili e una rappresentazione esplicita delle caratteristiche di interesse [3].

Con l'aumento della disponibilità di dati e l'evoluzione delle tecniche di deep learning, questo paradigma si è progressivamente spostato verso modelli capaci di apprendere rappresentazioni più generali dei dati [4, 5]. Un passaggio fondamentale in questa evoluzione è stato rappresentato dall'introduzione dell'architettura Transformer e del meccanismo di attention, originariamente sviluppato nell'ambito dell'elaborazione del linguaggio naturale [6]. Il meccanismo di attention permette al modello di attribuire un peso differente agli elementi di un insieme di dati in funzione delle loro relazioni e del loro contesto. L'idea di utilizzare il contesto (*in-context learning*) per ottenere una rappresentazione più informativa dell'input ha successivamente portato allo sviluppo di modelli in grado di svolgere compiti anche a partire da pochi esempi forniti direttamente nel contesto.

Parallelamente a questa evoluzione, il paradigma dei Foundation Models [7] ha iniziato a essere applicato anche a dati differenti dal linguaggio naturale. I Foundation Models sono modelli di grandi dimensioni pre-addestrati su dataset estesi e successivamente adattabili

a differenti compiti. Nel caso di dati tabellari e strutturati, come quelli clinici, sono stati sviluppati i cosiddetti Tabular Foundation Models (TFM) [8].

Un elemento particolarmente importante di questi modelli è la capacità di apprendere rappresentazioni latenti dei dati (embedding) che permettono di codificare informazioni e relazioni non necessariamente esplicite nella rappresentazione originale. Nel caso di dati genetici, un embedding mutazionale può quindi fornire una rappresentazione di una mutazione o di un profilo mutazionale, consentendo al modello di sfruttare informazioni sulle relazioni tra alterazioni genetiche che potrebbero non essere esplicitamente contenute nella semplice codifica binaria della loro presenza o assenza.

1.3 Obiettivo

L'obiettivo della presente tesi è lo sviluppo di una pipeline predittiva per l'identificazione del fenotipo fibrotico nelle neoplasie mieloproliferative. In questo contesto, il lavoro si articola in due principali linee di indagine. In primo luogo, si intende valutare nuovi modelli predittivi, i Tabular Foundation Models, in grado di identificare la fibrosi a partire da dati clinici e genetici. In secondo luogo, si vuole verificare se l'integrazione degli embeddings mutazionali, rappresentazioni latenti delle alterazioni genetiche più informative delle singole mutazioni genetiche, consenta di migliorare le prestazioni predittive dei modelli utilizzati, analizzandone l'effetto quando utilizzate come feature aggiuntiva nei Tabular Foundation Models.

La presente tesi si propone di sviluppare e valutare una pipeline predittiva per l'identificazione della diagnosi fibrotica nelle neoplasie mieloproliferative, integrando informazioni cliniche e genetiche mediante Tabular Foundation Models.

Capitolo 2

Background clinico

In questo capitolo verranno inizialmente descritte, in modo approfondito, le MPN e la loro classificazione secondo i criteri dell'Organizzazione Mondiale della Sanità (OMS), con particolare attenzione alla trombocitemia essenziale e alla mielofibrosi primaria in fase prefibrotica e fibrotica. Successivamente, saranno approfonditi i principali aspetti genetici, le manifestazioni cliniche e gli attuali metodi diagnostici.

2.1 Neoplasie Mieloproliferative

Le neoplasie mieloproliferative (MPN) sono un gruppo di tumori caratterizzate dalla proliferazione clonale di cellule staminali e progenitrici ematopoietiche della linea mieloide, responsabili della produzione di eritrociti, leucociti e piastrine. La proliferazione anomala di queste cellule può determinare un'eccessiva produzione di una o più linee cellulari del sangue e, nelle fasi più avanzate della malattia, alterare la normale morfologia e funzionalità del midollo osseo.

Esse comprendono la leucemia mieloide cronica (LMC), caratterizzata dalla presenza del riarrangiamento BCR::ABL1 che da origine al cromosoma Philadelphia (Ph+) [1] e le MPN

Ph− quali la trombocitemia essenziale (TE), la mielofibrosi (MF) e la policitemia vera (PV), che verranno descritte singolarmente nei successivi sottocapitoli.

2.2 Classificazione OMS

La classificazione dell'Organizzazione Mondiale della Sanità (OMS/WHO) costituisce lo standard internazionale per la diagnosi di queste patologie, integrando criteri morfologici, citogenetici, molecolari e clinici [9]. La classificazione WHO è stata aggiornata nel 2016 e nuovamente nel 2022 [10], con l'obiettivo di riflettere le scoperte recenti nella biologia molecolare delle neoplasie mieloidi. L'aggiornamento più recente include l'integrazione di mutazioni genetiche specifiche come criteri diagnostici fondamentali, in particolare per le neoplasie mieloproliferative Ph−.

Tabella 2.1: Classificazione delle principali neoplasie mieloproliferative secondo l'OMS.

Patologia	Caratteristiche principali	Alterazioni molecolari associate
Leucemia mieloide cronica (LMC)	Proliferazione clonale della linea mieloide con aumento dei granulociti	Gene di fusione *BCR-ABL1*
Policitemia vera (PV)	Aumento predominante della massa eritrocitaria e dell'ematocrito	Mutazione *JAK2 V617F* nel 95% circa dei casi
Trombocitemia essenziale (TE)	Trombocitosi persistente e proliferazione dei megacariociti	Mutazioni *JAK2*, *CALR* o *MPL*

(continua nella pagina successiva)

Patologia	Caratteristiche principali	Alterazioni molecolari associate
Mielofibrosi primaria (MFP)	Fibrosi midollare progressiva, splenomegalia e citopenie	Mutazioni *JAK2*, *CALR* o *MPL*
Leucemia neutrofila cronica (CNL)	Neutrofilia persistente e proliferazione granulocitaria	Frequenti mutazioni *CSF3R*
Leucemia eosinofila cronica (CEL)	Espansione clonale della linea eosinofila	Riarrangiamenti di *PDGFRA*, *PDGFRB* o *FGFR1*
MPN non classificabile (MPN-U)	Casi che non soddisfano pienamente i criteri diagnostici delle altre MPN	Variabile, spesso mutazioni driver delle MPN classiche

2.2.1 Trombocitemia essenziale

La trombocitemia essenziale è caratterizzata da trombocitosi persistente e significativa, con aumento del numero di megacariociti (cellule 10 a 15 volte più grandi di un normale globulo rosso, responsabili della produzione delle piastrine) nel midollo osseo. Essa rappresenta una delle più frequenti neoplasie mieloproliferative Ph−, con un'incidenza (1,8/100.000 persone-anno) generalmente superiore a quella della mielofibrosi primaria (0,5) e, in alcun casi, anche a quella della policitemia vera (1,7) [11].

La patogenesi molecolare della TE è legata a mutazioni in tre geni principali: *JAK2* V617F (variante del gene Janus Kinase caratterizzata dalla sostituzione dell'amminoacido valina con fenilalanina in posizione 617 della proteina, circa 50-60% dei casi), *CALR* (calreticulina, circa 20-30% dei casi) e *MPL* (recettore della trombopoietina, circa 3-5% dei casi). Queste mutazioni attivano la via di segnalazione JAK-STAT [12, 13], promuovendo la proliferazione

megacariocitaria.

2.2.2 Mielofibrosi

La mielofibrosi è una rara neoplasia mieloproliferativa cronica del midollo osseo che può essere primaria, se insorge autonomamente, o secondaria, se evolve da precedenti patologie come la policitemia vera o la trombocitemia essenziale. La mielofibrosi si caratterizza per una fibrosi progressiva del midollo osseo, che ostacola la produzione normale dei globuli; essa può presentarsi in due forme: fase pre-fibrotica (con proliferazione cellulare senza fibrosi significativa) e fase fibrotica (con fibrosi midollare marcata).

Per compensare, il corpo cerca di produrre cellule sanguigne fuori dal midollo, soprattutto nel fegato e nella milza, causando un'emopoiesi extramidollare. Presenta un quadro clinico variable, che può includere anemia, leucocitosi, trombocitosi e splenomegalia marcata. Essa rappresenta la forma più aggressiva di MPN Ph− [14, 15].

Le mutazioni molecolari caratteristiche sono identiche a quelle della PV e della TE (*JAK2*, *CALR*, *MPL*), presenti in circa 85-90% dei casi. La presenza di mutazioni aggiuntive in geni come *ASXL1*, *EZH2*, *IDH1/2*, e *SRSF2* è associata a prognosi peggiore.

2.3 Aspetti genetici e mutazioni driver

Le neoplasie mieloproliferative Ph− sono caratterizzate dalla presenza di alterazioni genetiche acquisite appartenenti alle cellule staminali ematopoietiche, che conferiscono loro un vantaggio proliferativo.

Durante la vita della cellula, essa può acquisire numerose alterazioni genetiche ma non tutte hanno la stessa importanza. Alcune mutazioni sono semplicemente presenti nel genoma delle cellule, ma non modificano in maniera significativa il loro comportamento. Queste vengono generalmente chiamate mutazioni passenger. Le mutazioni driver, invece, modificano il com-

portamento della cellula e le conferiscono un vantaggio selettivo, per esempio aumentando la capacità di proliferare o di sfuggire ai normali meccanismi di regolazione.

Nelle MPN Ph−, le principali mutazioni driver interessano i geni *JAK2*, *CALR* e *MPL*. Sebbene si tratti di alterazioni differenti, esse convergono sull'attivazione anomala della via di segnalazione JAK-STAT, un sistema fondamentale per la regolazione della proliferazione, della differenziazione e della sopravvivenza delle cellule ematopoietiche. Pertanto, la loro scoperta ha rivoluzionato la comprensione della patogenesi di queste malattie e ha contribuito a migliorarne la classificazione diagnostica.

In condizioni fisiologiche, le cellule staminali ematopoietiche non proliferavano casualmente, ma la loro crescita e differenziazione è regolata da numerosi fattori di crescita, che trasmettono segnali alle cellule ematopoietiche legandosi a specifici recettori presenti sulla loro superficie.

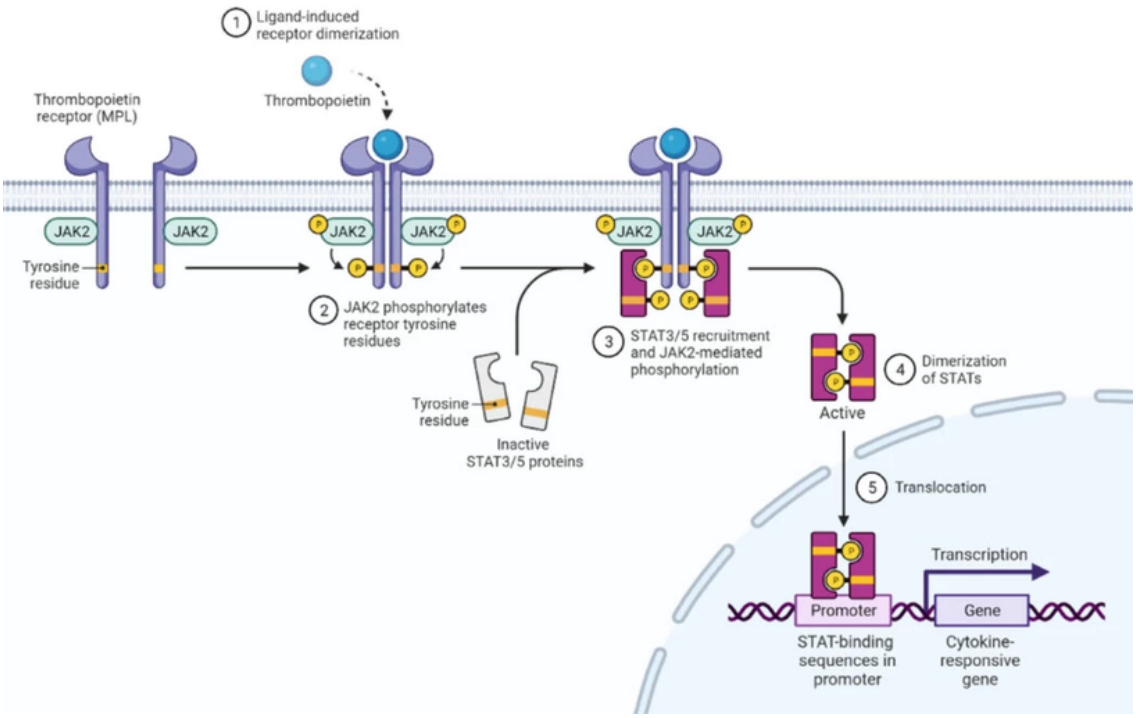

Figura 2.1: Via di segnalazione JAK/STAT.

Questi segnali vengono trasmessi dall'esterno all'interno della cellula attraverso specifiche vie di segnalazione intracellulari: una delle più importanti è la via JAK-STAT (*Signal Transducers and Activators of Transcription*) [12, 13].

Il legame del fattore di crescita al recettore determina l'attivazione delle proteine della famiglia delle Janus Kinase (JAK), tra cui JAK2 [16, 13]. Le JAK attivate fosforilano quindi le proteine della famiglia STAT (*Signal Transducers and Activators of Transcription*), che vengono a loro volta attivate e traslocano nel nucleo. Una volta raggiunto il nucleo, le proteine STAT regolano l'espressione di geni coinvolti nella proliferazione, nella differenziazione e nella sopravvivenza cellulare.

Nelle MPN, alterazioni a carico dei componenti di questa via possono determinare una sua attivazione persistente e non adeguatamente regolata, favorendo l'espansione delle cellule ematopoietiche che presentano alterazioni del DNA. La mutazione più frequente è *JAK2* V617F, identificata in circa il 95% dei pazienti con policitemia vera e nel 50–60% dei soggetti con trombocitemia essenziale o mielofibrosi primaria.

Nei pazienti privi della mutazione *JAK2*, possono essere presenti mutazioni nei geni *CALR* e *MPL*. Le mutazioni di *CALR* sono particolarmente frequenti nella trombocitemia essenziale e nella mielofibrosi primaria e risultano generalmente associate a una prognosi più favorevole rispetto a *JAK2*. Le mutazioni di *MPL*, meno frequenti, interessano il recettore della trombopoietina e contribuiscono anch'esse all'attivazione delle vie proliferative [17, 18].

Oltre alle mutazioni driver, possono essere presenti alterazioni nei geni coinvolti nella regolazione epigenetica, nello splicing dell'RNA, nella risposta al danno del DNA e nella progressione leucemica. Tra questi ci sono: *ASXL1*, *TET2*, *DNMT3A*, *EZH2*, *SRSF2*, *U2AF1* e *TP53*. Di conseguenza, il profilo mutazionale può contribuire alla classificazione della fibrosi e alla stratificazione prognostica [19].

2.4 Manifestazioni cliniche

Le manifestazioni cliniche delle MPN derivano sia dall'aumento della produzione cellulare sia dall'attivazione infiammatoria e dalla disfunzione delle cellule ematiche.
Le complicanze più importanti comprendono: eventi trombotici arteriosi o venosi, episodi emorragici, splenomegalia ed epatomegalia (espressione di emopoiesi extramidollare) e alterazioni della conta eritrocitaria, leucocitaria o piastrinica [20].

L'emocromo nella mielofibrosi è variabile e può comprendere nelle fasi iniziali la leucocitosi e la trombocitosi.
Nelle fasi più avanzate, la fibrosi midollare può causare pancitopenia, ovvero la diminuzione di tutte le cellule del sangue e quindi, al contrario, possono prevalere leucopenia e piastrinopenia [21].

Inoltre, tutte le MPN sono accumunate da un incrementato rischio di evoluzione in leucemia mieloide acuta (LMA).

2.5 Diagnosi

La diagnosi delle neoplasie mieloproliferative si basa sull'integrazione di dati clinici, ematologici, morfologici e molecolari. Secondo i criteri diagnostici dell'Organizzazione Mondiale della Sanità (OMS), la diagnosi non dipende infatti dalla presenza di un singolo parametro, ma dalla combinazione di specifici criteri maggiori e minori, che variano in funzione dell'entità patologica considerata.

La valutazione iniziale comprende l'esame emocromocitometrico, che consente di evidenziare alterazioni quali eritrocitosi, trombocitosi o leucocitosi. Tali risultati devono essere interpretati nel contesto clinico del paziente e integrati con ulteriori indagini.

Nei pazienti con sospetta MPN, si prosegue con la ricerca delle principali mutazioni driver delle MPN Ph−. Generalmente, è ricercata inizialmente la mutazione *JAK2 V617F* e in caso

di negatività, l'analisi può essere estesa alle mutazioni driver dei geni *CALR* e *MPL*. L'assenza di tutte e tre le principali mutazioni driver, definita condizione "triplo negativa", non esclude comunque la presenza di una MPN e rende necessaria una valutazione complessiva degli altri criteri diagnostici, come la valutazione istologica del midollo.

La biopsia osteomidollare rappresenta un passaggio fondamentale nel percorso diagnostico, poiché permette di valutare la cellularità del midollo osseo, la morfologia dei megacariociti e il grado di fibrosi. Quest'ultimo parametro assume particolare rilevanza nella distinzione tra trombocitemia essenziale, mielofibrosi prefibrotica e mielofibrosi fibrotica [10, 9]. La corretta classificazione diagnostica è importante, poiché differenti entità cliniche possono presentare caratteristiche sovrapponibili nelle fasi iniziali della malattia.

Tabella 2.2: Gradi di fibrosi midollare.

Grado	**Descrizione**
0	Assenza di fibrosi significativa.
1	Fibrosi reticolinica lieve.
2	Fibrosi reticolinica aumentata e diffusa.
3	Fibrosi reticolinica molto densa e diffusa.

Ulteriori informazioni possono essere ottenute attraverso il sequenziamento mediante Next Generation Sequencing (NGS), una tecnologia che consente di analizzare contemporaneamente numerose regioni del DNA. A partire dal campione biologico del paziente, il DNA viene estratto e frammentato in numerose sequenze più piccole, che vengono successivamente amplificate e sequenziate in parallelo. Le sequenze ottenute vengono quindi ricostruite e confrontate con una sequenza di riferimento, permettendo di identificare eventuali varianti genetiche presenti nel campione.
L'analisi NGS può comprendere numerosi geni e consentire quindi di identificare, oltre alle principali mutazioni driver, anche alterazioni aggiuntive, fornendo un profilo mutazionale somatico del paziente potenzialmente rilevante per la caratterizzazione della malattia e, in alcuni casi, per la valutazione prognostica.

Capitolo 3

Background metodologico

In questo capitolo verranno introdotti i principali concetti e strumenti metodologici alla base del presente lavoro. Verranno inizialmente descritti gli approcci di representation learning e il concetto di embedding, con riferimento alla loro applicazione alla rappresentazione dei dati genetici. Infine, verranno presentati i Tabular Foundation Models e i principi alla base del loro funzionamento.

3.1 Analisi esplorativa genomica

L'analisi esplorativa dei dati genomici costituisce una fase preliminare del lavoro, finalizzata a caratterizzare la distribuzione delle alterazioni genetiche e a identificare eventuali relazioni statistiche presenti nei dati. Tale analisi consente di valutare la frequenza delle alterazioni, le associazioni tra geni e la struttura complessiva dello spazio genomico, fornendo una prima descrizione delle caratteristiche dei dati che saranno successivamente utilizzati per la costruzione delle rappresentazioni mediante embedding.

3.1.1 Dati genomici

I dati genomici ottenuti mediante tecniche di Next Generation Sequencing (NGS) sono rappresentati attraverso l'insieme delle alterazioni genetiche identificate nei singoli pazienti. Una rappresentazione comune consiste nella costruzione di una matrice *paziente* $\times$ *gene*, nella quale ciascuna riga rappresenta un paziente e ciascuna colonna un gene, mentre il valore associato a ciascuna coppia paziente-gene indica la presenza o l'assenza di una mutazione associata al gene.

Tale rappresentazione consente di descrivere direttamente il profilo mutazionale dei pazienti, ma presenta alcune caratteristiche che devono essere considerate nell'analisi. I dati genomici sono caratterizzati da elevata dimensionalità e sparsità, poiché ciascun paziente presenta generalmente alterazioni in una frazione limitata dei geni considerati. Inoltre, la rappresentazione convenzionale tratta ciascun gene come una variabile distinta e non incorpora esplicitamente eventuali relazioni tra geni.

Queste caratteristiche rendono utile una fase preliminare di caratterizzazione dei dati e motivano successivamente l'utilizzo di embeddings, in grado di sintetizzare l'informazione genomica in uno spazio continuo.

3.1.2 Frequenza mutazionale

Una prima modalità per caratterizzare la distribuzione delle alterazioni genomiche presenti nel dataset consiste nel calcolare la frequenza mutazionale [22], una misura utilizzata per quantificare la proporzione di pazienti nei quali un determinato gene presenta un'alterazione. Per ogni gene, è definita come:

$$f_j = \frac{n_j}{N} \tag{3.1}$$

dove n_j rappresenta il numero di pazienti portatori di una mutazione nel gene j e N il numero totale di pazienti analizzati.

L'analisi delle frequenze permette di identificare i geni maggiormente alterati nella popolazione studiata e di evidenziare potenziali eventi biologicamente rilevanti. Le frequenze mutazionali sono state successivamente rappresentate mediante grafici a barre impilate, stratificate per gene driver e per diagnosi (sottosezione 5.2.1).

3.1.3 Correlazioni tra geni

Le alterazioni genomiche presenti nei pazienti non devono necessariamente essere considerate indipendenti. L'analisi delle associazioni tra geni permette infatti di identificare pattern di co-occorrenza o di mutua esclusività tra differenti alterazioni.

A partire dalla matrice paziente-gene, è possibile costruire una matrice di associazione nella quale ciascun elemento rappresenta il grado di relazione tra due geni. La scelta della misura di associazione deve essere effettuata tenendo conto della natura dei dati, che nel caso di una rappresentazione binaria descrivono la presenza o l'assenza di un'alterazione.

Per visualizzare le relazioni tra i diversi geni è stata costruita una matrice di associazione, successivamente rappresentata mediante heatmap con clustering gerarchico (sottosezione 5.2.2). L'associazione tra le coppie di geni è stata valutata mediante test esatto di Fisher [23] e quantificata attraverso l'odds ratio (OR), definito come:

$$OR = \frac{a \cdot d}{b \cdot c} \tag{3.2}$$

dove a, b, c e d rappresentano le frequenze osservate nella tabella di contingenza.

	Gene B presente	**Gene B assente**
Gene A presente	a	b
Gene A assente	c	d

Tabella 3.1: Tabella di contingenza per l'analisi dell'associazione tra Gene A e Gene B

Poiché l'odds ratio può assumere valori compresi tra 0 e infinito, per una migliore rappresen-

tazione grafica è utilizzata la trasformazione logaritmica in base 2:

$$\log_2(OR) \tag{3.3}$$

Per tenere conto del problema dei confronti multipli, i valori di p ottenuti mediante test esatto di Fisher sono corretti utilizzando la procedura di Benjamini-Hochberg [24] per il controllo del False Discovery Rate (FDR), ossia della proporzione attesa di falsi positivi tra i risultati identificati come significativi.

La procedura prevede l'ordinamento crescente dei valori di p ottenuti dai singoli test e la loro correzione in funzione del numero complessivo di confronti e della posizione occupata da ciascun valore nell'ordinamento. In questo modo è possibile ridurre l'effetto dei confronti multipli mantenendo una maggiore capacità di individuare associazioni significative rispetto a correzioni più conservative.

3.1.4 Riduzione dimensionale e visualizzazione

L'elevata dimensionalità dei dati genomici rende utile l'applicazione di tecniche di riduzione dimensionale finalizzate all'esplorazione della struttura latente della popolazione studiata.

Tra le principali tecniche [25] impiegate figurano:

- Principal Component Analysis (PCA): tecnica lineare che proietta i dati lungo direzioni ortogonali, dette *componenti principali*, individuate in modo da massimizzare progressivamente la varianza spiegata;
- t-distributed Stochastic Neighbor Embedding (t-SNE): tecnica non lineare progettata per preservare principalmente le relazioni locali tra le osservazioni;
- Uniform Manifold Approximation and Projection (UMAP): tecnica non lineare che costruisce una rappresentazione a bassa dimensionalità preservando sia le strutture locali sia, in parte, quelle globali dei dati. Rispetto a t-SNE, UMAP tende a mantenere meglio la struttura complessiva della distribuzione.

Tali metodologie consentono di rappresentare ciascun paziente mediante coordinate bidimensionali o tridimensionali, facilitando l'identificazione di cluster naturali e sottogruppi molecolari.

Le rappresentazioni ottenute vengono utilizzate per:

- identificare eventuali sottotipi genomici;
- visualizzare la distribuzione delle mutazioni driver;
- esplorare relazioni tra profili genomici e caratteristiche cliniche;
- evidenziare possibili gruppi di pazienti con caratteristiche biologiche condivise.

3.2 Costruzione degli embeddings

Il concetto di embedding ha assunto particolare rilevanza nell'ambito dell'elaborazione del Natural Language Processing. In questo contesto, una parola può essere rappresentata attraverso un vettore numerico appreso a partire dai contesti linguistici nei quali essa compare. L'idea alla base di questi approcci è che il significato di una parola possa essere descritto, almeno in parte, attraverso le altre parole con cui essa tende a comparire. Parole che presentano contesti simili possono quindi assumere rappresentazioni vettoriali simili.

3.2.1 Stato dell'arte

Tra i metodi che hanno contribuito maggiormente alla diffusione degli embedding nel campo della NLP vi è `Word2Vec` introdotto da Mikolov et al. [26]. `Word2Vec` permette di apprendere rappresentazioni vettoriali delle parole a partire da grandi quantità di testo, sfruttando le informazioni relative al contesto in cui le parole compaiono.

La sua architettura comprende due modelli: Continuous Bag-of-Words (CBOW) e Skip-Gram. Nel modello CBOW l'obiettivo è prevedere una parola a partire dalle parole che la

circondano, mentre nel modello Skip-Gram il processo viene invertito: data una parola, il modello cerca di prevedere le parole che compaiono nel suo contesto.

Se si considera una frase costituita da una sequenza di parole e una determinata parola viene scelta come elemento centrale, le parole che si trovano all'interno di una certa finestra intorno ad essa costituiscono il suo contesto. Il modello Skip-Gram utilizza queste coppie parola-contesto durante l'addestramento e modifica progressivamente i parametri della rete in modo da aumentare la probabilità di prevedere correttamente le parole che compaiono nel contesto dell'elemento considerato. La rappresentazione vettoriale appresa nella componente interna del modello costituisce quindi l'embedding della parola.

Durante l'addestramento, parole che compaiono frequentemente in contesti simili ricevono rappresentazioni vettoriali che tendono a essere simili. In questo modo, le relazioni apprese dal modello non vengono espresse attraverso una singola variabile binaria, ma distribuite tra le diverse componenti del vettore. La rappresentazione risultante può quindi contenere informazioni sulle relazioni tra gli elementi che non erano esplicitamente presenti nella codifica iniziale.

Una caratteristica importante di Skip-Gram è che il modello non richiede di definire necessariamente a priori quali siano le relazioni semantiche tra le parole. Queste relazioni vengono apprese direttamente dai pattern di co-occorrenza osservati nel corpus.

Il principio di Skip-Gram è stato successivamente adattato alla rappresentazione di dati genomici nel lavoro di Kim et al. [27] che descrive il modello **Mut2vec**.

3.2.2 Gene-level embedding (Mut2vec)

Il principio alla base degli embedding può essere trasferito, infatti, a domini differenti dal linguaggio naturale. L'analogia tra linguaggio e dati biologici può essere estesa anche ai profili mutazionali dei tumori. In questo caso, invece di considerare una frase composta da parole, è possibile considerare ciascun paziente come un insieme di alterazioni genetiche. I geni mutati nello stesso paziente possono quindi essere considerati come elementi che condividono un

determinato contesto. In questa prospettiva, la co-occorrenza di alterazioni genetiche può essere utilizzata per apprendere rappresentazioni vettoriali dei geni o delle mutazioni.

Il modello `Mut2Vec` rappresenta un'evoluzione importante rispetto alla semplice rappresentazione binaria dei dati genomici. Una matrice paziente-gene codificata mediante valori 0 e 1 indica infatti esclusivamente se una determinata alterazione è presente o assente, senza rappresentare direttamente le relazioni tra le diverse alterazioni. Gli embedding appresi da `Mut2Vec` permettono, invece, di rappresentare ogni gene mediante un vettore 300D [27] e di utilizzare la posizione relativa dei vettori nello spazio latente come informazione sulle relazioni apprese dai dati.

Tabella 3.2: Corrispondenza tra i principali concetti del Natural Language Processing (NLP) e la loro applicazione alla rappresentazione dei dati genomici.

NLP	Genomica
Documento	Paziente
Parola	Gene
Contesto della parola	Altri geni mutati nello stesso paziente
Word2Vec	Mut2Vec
Word embedding	Gene embedding

Un ulteriore aspetto del modello riguarda l'integrazione di informazioni biologiche esterne. Gli autori evidenziano infatti che i dati mutazionali disponibili possono essere relativamente limitati e che le mutazioni rare possono fornire poche informazioni utili per l'apprendimento basato sulla sola co-occorrenza. Per affrontare questo problema, `Mut2Vec` integra informazioni provenienti dalla letteratura biomedica e dalle reti di interazione proteina-proteina, utilizzando anche una procedura di *retrofitting* per modificare gli embedding sulla base delle relazioni biologiche note [27].

Tuttavia, il modello non distingue direttamente tra diverse varianti dello stesso gene, ad

esempio una mutazione missenso o una delezione; questo significa che tutte le mutazioni appartenenti allo stesso gene condividono la stessa rappresentazione vettoriale. Questa scelta rende il modello più robusto dal punto di vista statistico, dato che molte varianti specifiche sono rare, ma comporta la perdita di informazioni dettagliate sull'effetto funzionale delle singole mutazioni.

Gli autori rilasciano i vettori di embedding pre-trained per ciascun gene (identificato con nomenclatura Ensemble Gene ID) di dimensione 300D, ottenuti da diversi profili mutazionali di pazienti oncologici e non da un singolo tipo di cancro (circa 13.000 campioni tumorali del database International Cancer Genome Consortium [28]).

3.2.3 Patient-level embedding

Oniani et al. [29] estendono il concetto di gene-level embedding alla rappresentazione a livello di paziente (*patient-level embedding*).

In particolare, gli embedding genetici ottenuti da `Mut2Vec` vengono utilizzati come rappresentazioni latenti dei geni mutati e, successivamente, per ciascun paziente, viene costruito un vettore rappresentativo calcolando la media degli embedding associati alle mutazioni presenti nel profilo genomico del paziente.

$$AE(p) = \frac{1}{|pv_p|} \sum_{me_p \in pv_p} f_c(me_p) \tag{3.4}$$

dove:

- p indica il paziente considerato;
- pv_p rappresenta l'insieme delle feature associate al paziente p;
- $|pv_p|$ indica il numero di feature presenti nell'insieme pv_p;
- me_p rappresenta una singola mutazione o feature appartenente all'insieme pv_p;

- $f_c(me_p)$ indica l'embedding associato alla mutazione me_p;
- $AE(p)$ rappresenta l'*average embedding* del paziente p, ottenuto calcolando la media degli embedding delle feature presenti nel suo profilo.

Un altro metodo di aggregazione è l'attention pooling [30]: rappresenta una strategia avanzata per la costruzione dell'embedding del paziente, nella quale il contributo delle singole mutazioni viene appreso automaticamente dal modello durante la fase di training.

Diversamente dal mean pooling, in cui tutti i geni contribuiscono in egual misura alla rappresentazione finale, e weighted pooling, in cui i pesi α_i vengono definiti a priori sulla base della frequenza mutazionale o dell'importanza biologica, l'attention pooling consente di assegnare pesi differenti alle mutazioni in maniera dinamica e dipendente dal contesto biologico.

$$z_p = \sum_i \alpha_i e_i \tag{3.5}$$

dove:

- e_i rappresenta l'embedding del gene mutato i;
- α_i rappresenta il peso associato al gene i.

I coefficienti di attenzione vengono appresi automaticamente mediante una funzione parametrica seguita da una normalizzazione softmax:

$$\alpha_i = \text{softmax}(f(e_i))$$

dove $f(\cdot)$ rappresenta una funzione neurale apprendibile che stima la rilevanza relativa di ciascun gene mutato.

In questo modo, mutazioni maggiormente informative ricevono pesi più elevati, mentre alterazioni meno rilevanti contribuiscono in misura minore alla rappresentazione finale del paziente. Tale approccio consente di modellare relazioni biologiche complesse, quali cooperazione clonale e combinazioni specifiche di alterazioni genetiche.

Questo aspetto risulta particolarmente rilevante nell'ambito delle neoplasie mieloproliferative, in cui il significato biologico e prognostico di una mutazione può dipendere dalla presenza di specifiche co-mutazioni. Ad esempio, una mutazione in $TET2$ può assumere un ruolo differente se associata a $JAK2$, $SRSF2$ oppure $ASXL1$. Analogamente, mutazioni ad alto impatto prognostico quali $TP53$, $EZH2$ o $ASXL1$ possono contribuire in misura diversa alla definizione dello stato patologico del paziente.

Tuttavia, l'aumento della complessità del modello richiede strategie di regolarizzazione appropriate, al fine di ridurre il rischio di overfitting nelle coorti cliniche di dimensioni limitate.

3.3 Tabular Foundation Models

I Tabular Foundation Models (TFM) rappresentano una recente evoluzione nell'ambito dell'apprendimento automatico per dati tabellari. Ispirati al paradigma dei Foundation Models sviluppato per il linguaggio naturale, questi modelli vengono pre-addestrati su milioni di dataset eterogenei sintetici, con l'obiettivo di apprendere rappresentazioni trasferibili a nuovi problemi predittivi. Grazie alla loro capacità di sfruttare conoscenza acquisita durante il pretraining, i TFM hanno dimostrato prestazioni competitive in diversi compiti di classificazione e regressione, anche in presenza di dataset di dimensioni limitate.

3.3.1 Principio di funzionamento e tipologie

I Tabular Foundation Models si basano sull'architettura Transformer (simile ai modelli linguistici), originariamente introdotta per l'elaborazione del linguaggio naturale. L'idea alla base del Transformer è quella di utilizzare il meccanismo di **self-attention** per permettere al modello di valutare le relazioni tra i diversi elementi presenti nell'input [6].

Il meccanismo di self-attention si basa sulla costruzione di tre rappresentazioni, denominate *Query*, *Key* e *Value*. Dato un input rappresentato dalla matrice X, queste vengono ottenute attraverso trasformazioni lineari:

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \tag{3.6}$$

dove W_Q, W_K e W_V sono matrici di parametri apprese durante l'addestramento. L'attenzione viene quindi calcolata confrontando ogni *Query* con le *Key* degli altri elementi:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.7}$$

dove d_k rappresenta la dimensionalità delle *Key*. Il prodotto QK^T misura la similarità tra gli elementi, mentre la funzione *softmax* trasforma tali valori in pesi che indicano quanto ciascun elemento debba contribuire alla rappresentazione finale. In questo modo, ogni elemento dell'input può essere rappresentato tenendo conto del contesto costituito dagli altri elementi presenti nell'input [6].

Come mostrato nella Tabella 3.3, i TFM possono essere classificati in differenti paradigmi in base alla strategia di apprendimento adottata. Tra questi, il modello TabPFN appartiene alla categoria delle **Prior-Fitted Networks**, mentre altri modelli recenti sfruttano approcci basati su in-context learning, pre-training auto-supervisionato o rappresentazioni semantiche delle feature tabellari.

Tabella 3.3: Principali paradigmi dei Tabular Foundation Models.

Paradigma	Modelli	Approccio	Caratteristica distintiva
ICL (In-Context Learning)	TabICL, OrionMSP, OrionBix, Mitra	Utilizza gli esempi del training set come contesto durante l'inferenza	Apprendimento senza aggiornamento dei parametri del modello (zero-shot o few-shot)

(continua alla pagina successiva)

Paradigma	Modelli	Approccio	Caratteristica distintiva
Prior-Fitted Network	TabPFN	Approssima l'inferenza bayesiana mediante una rete neurale pre-addestrata	Elevata efficienza su dataset di piccole e medie dimensioni e buona calibrazione delle probabilità
Denoising Transformer	TabDPT	Pre-training auto-supervisionato basato sulla ricostruzione di dati corrotti	Miglior compromesso tra prestazioni ed efficienza nel fine-tuning
Semantics-Aware ICL	ContextTab	Utilizza embedding semantici per rappresentare colonne e variabili categoriche	Sfrutta informazioni semantiche associate ai nomi delle feature e ai valori categorici
Probabilistic / ICL	LimiX	Modellazione probabilistica basata sulla stima della likelihood e mixture modeling	Predizioni accompagnate da una stima esplicita dell'incertezza

Questo panorama eterogeneo ha motivato l'estensione dell'analisi sperimentale a diversi modelli TFM, implementati nella libreria `TabTune` [31].

TabPFN

TabPFN (Tabular Prior-data Fitted Network) è il primo foundation model dedicato a dati tabulari caratterizzati da eterogeneità nelle tipologie e nelle distribuzioni delle feature, ed è progettato per fare *in-context learning*.

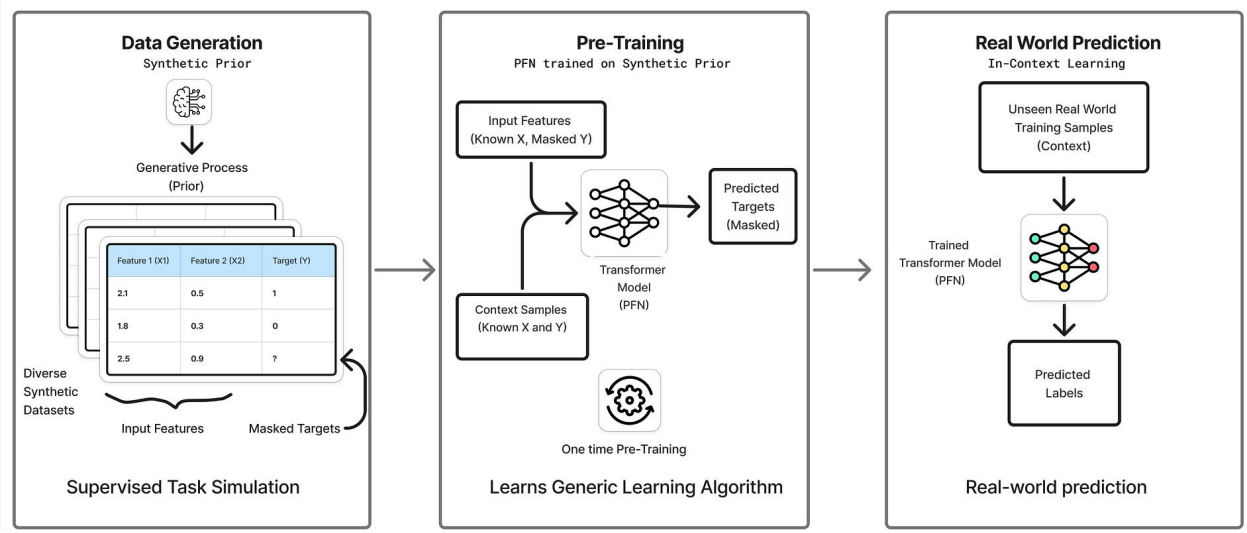

Figura 3.1: Schema di funzionamento del modello TabPFN [32].

Invece di ottimizzare i pesi per ogni dataset, durante il pretraining TabPFN viene preaddestrato su un grande numero di dataset sintetici e cerca di imparare una distribuzione prior (distribuzione di funzioni e relazioni feature-target) [33]. Possiamo indicare questi dataset come:

$$D^{(1)}, D^{(2)}, D^{(3)}, \ldots, D^{(N)}.$$

Ogni dataset sintetico può essere rappresentato come:

$$D^{(k)} = \left\{ (x_1^{(k)}, y_1^{(k)}), \ldots, (x_{n_k}^{(k)}, y_{n_k}^{(k)}) \right\} \tag{3.8}$$

Il modello cerca quindi di apprendere una funzione:

$$f_\theta(x, D) \approx p(y \mid x, D). \tag{3.9}$$

dove θ rappresenta i parametri appresi dal Transformer.

Terminata la fase di pretraining, il modello può essere applicato a un nuovo dataset .

$$D_{\text{new}} = \{(x_1, y_1), \ldots, (x_n, y_n)\}\,. \tag{3.10}$$

Supponiamo di avere una nuova osservazione x_*. Il modello stima il relativo target y_* usando, come contesto (*in-context learning*), le informazioni contenute nel dataset D_{new}.

$$p(y_* \mid x_*, D_{\text{new}}).$$

Quindi, in TabPFN il training set non viene utilizzato per aggiornare i pesi del modello (come nel machine learning tradizionale), ma fornisce l'informazione necessaria per effettuare la predizione durante l'inferenza su nuovi dati, in tempi brevissimi [33]. Infatti, i parametri θ del Transformer sono appresi durante il pretraining, pertanto non dovrà effettuare un nuovo training dei suoi parametri sul nuovo dataset.

$$\boxed{(D_{\text{new}}, x_*) \xrightarrow{f_\theta} p(y_* \mid x_*, D_{\text{new}})}.$$

TabICL

Il modello TabICL, a differenza di TabPFN, ha un'architettura pensata per essere più veloce e scalabile su dataset grandi, caratterizzati da features ad elevata dimensionalità e da relazioni potenzialmente complesse. La sua architettura, detta *two-stage* [34], è costituita da:

1. **Column-then-row attention**
    - In una prima fase, un Transformer *column-wise* costruisce una rappresentazione latente delle singole feature, tenendo conto della distribuzione dei valori lungo la colonna (quindi sensibile a scale, sparsità, ecc.)
    - Successivamente, un Transformer *row-wise* combina le rappresentazioni delle diverse feature e modella le loro interazioni all'interno della singola osservazione, ottenendo una rappresentazione a dimensionalità fissa per ciascuna riga.
2. **Dataset-wise In-Context Learning**
    - Le rappresentazioni ottenute per le osservazioni di training e di test vengono successivamente fornite a un ulteriore Transformer, che opera a livello di dataset.
    - Le osservazioni del training set, associate alle rispettive etichette, costituiscono il contesto utilizzato dal modello per effettuare la predizione delle etichette delle osservazioni di test.

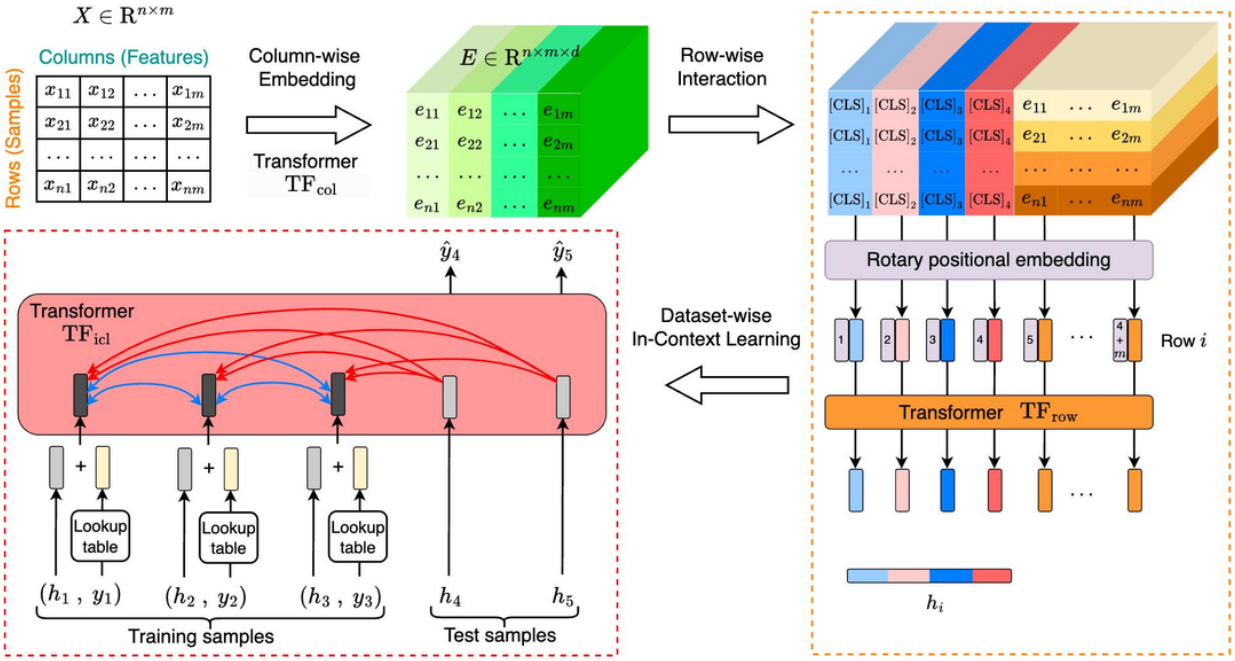

Figura 3.2: Architettura del modello TabICL [34].

Questa struttura permette di evitare alcuni dei costi dell'attenzione row-by-row di TabPFN.

Mitra

Un aspetto rilevante nell'impiego dei TFM riguarda la costruzione dei *prior* sintetici utilizzati durante il pre-training, dai quali dipende in parte la capacità dei modelli di apprendere strutture e relazioni generalizzabili ai dati reali.

Zhang et al. [35] analizzano in modo sistematico le caratteristiche dei *prior* sintetici, mettendo in evidenza come la loro composizione e complessità possano influenzare significativamente le prestazioni dei TFMs. Il lavoro studia quali proprietà dei *prior*, tra cui diversità, distintività e capacità di generalizzazione a dati reali, risultino maggiormente associate a migliori performance, proponendo un mix curato di *prior* sintetici.

Pertanto, l'obiettivo di Mitra non è fornire una nuova architettura, ma ampliare la varietà dei prior sintetici utilizzati durante il pre-training, fornendo una prior mista capace di generalizzare meglio su dataset reali.

3.3.2 La libreria `TabTune`

La libreria open-source `TabTune` [31], è stata sviluppata con l'obiettivo di fornire un framework unificato per l'applicazione, il confronto e l'adattamento di differenti Tabular Foundation Models. La libreria mette a disposizione un'interfaccia omogenea che semplifica le fasi di preprocessing, addestramento, ottimizzazione degli iperparametri e valutazione delle prestazioni, consentendo di confrontare modelli eterogenei mediante una pipeline comune.

Uno dei principali vantaggi di `TabTune` è la possibilità di utilizzare differenti categorie di TFM attraverso un'unica API, rendendo più agevole l'analisi comparativa tra diverse architetture e strategie di tuning. La libreria integra inoltre procedure automatiche per la gestione delle variabili numeriche e categoriche, garantendo una maggiore standardizzazione degli esperimenti.

Un ulteriore aspetto rilevante è il supporto a diverse strategie di adattamento dei modelli al dataset target. In particolare, `TabTune` consente sia di utilizzare i modelli direttamente in modalità di inferenza, sfruttando esclusivamente la conoscenza acquisita durante il pre-training, sia di applicare procedure di fine-tuning complete o efficienti dal punto di vista computazionale. Questa flessibilità permette di valutare il contributo del pre-training e di analizzare il comportamento dei modelli in differenti scenari sperimentali.

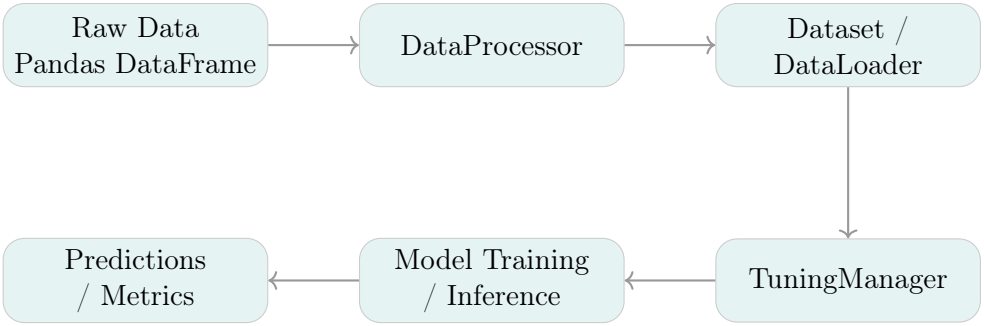

Figura 3.3: Overview del framework TabTune.

3.3.3 Strategie di adattamento dei pesi

I Tabular Foundation Models possono essere impiegati su nuovi dataset target secondo diverse strategie di adattamento, che differiscono per grado di aggiornamento dei parametri, costo computazionale e quantità di dati richiesti.

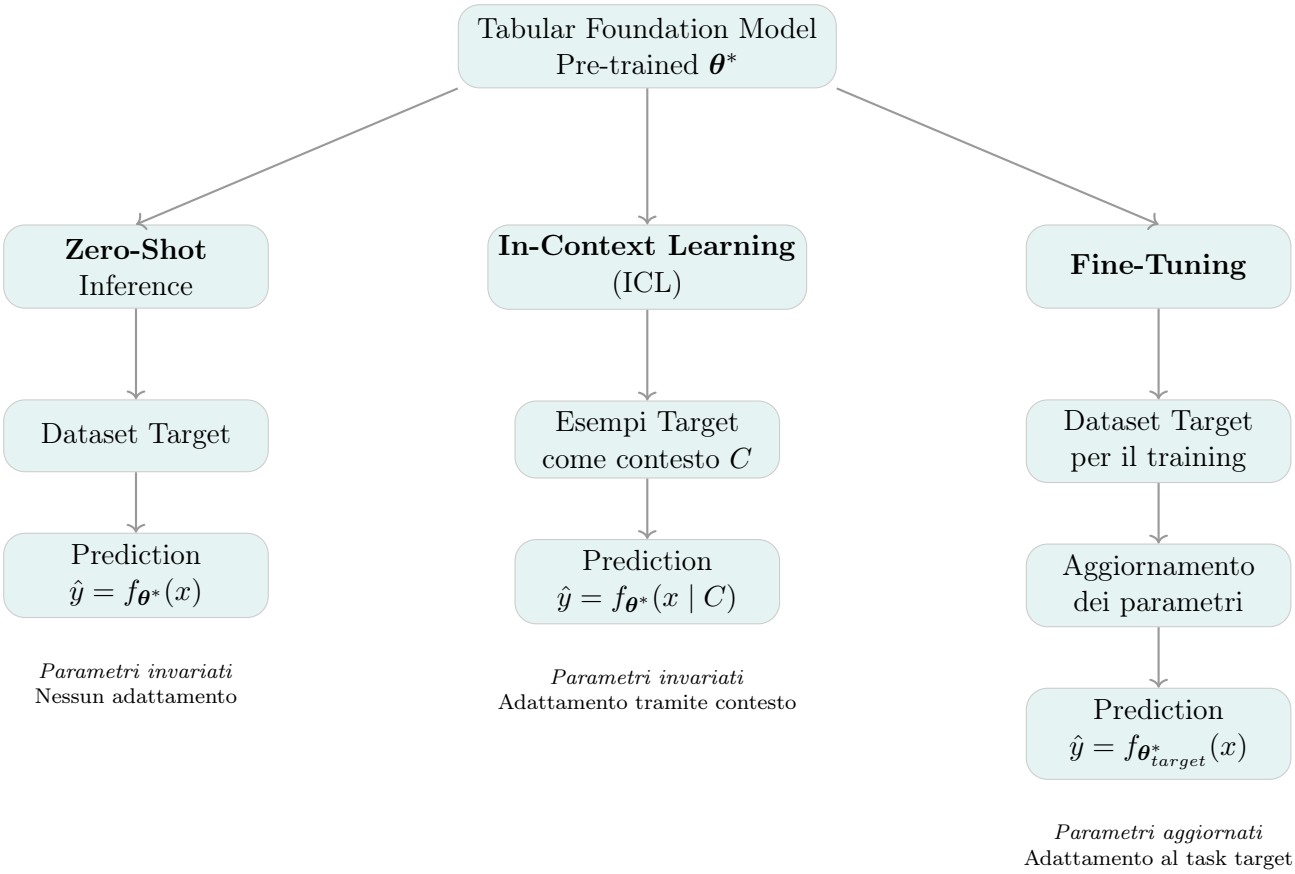

Figura 3.4: Principali paradigmi di adattamento di un Foundation Model.

In generale, è possibile distinguere tre modalità principali:

1. **Zero-shot inference**. Il termine *zero-shot* indica l'assenza di esempi del task target utilizzati per modificare il modello prima della predizione. Il dataset target viene impiegato esclusivamente per generare le predizioni e non interviene nel processo di ottimizzazione dei parametri del modello: tutta l'informazione necessaria al modello deriva dalla fase di pre-training.

Il principale vantaggio dello Zero-Shot è quindi rappresentato dalla semplicità e dal basso costo computazionale [31]. Esso risulta particolarmente interessante quando il dataset target contiene un numero limitato di osservazioni oppure quando si desidera misurare la capacità intrinseca di generalizzazione del Foundation Model. D'altra parte, l'assenza di adattamento può limitare le prestazioni quando il dataset target presenta caratteristiche significativamente differenti da quelle osservate durante il pre-training.

2. **In-context learning**. E' un paradigma intermedio tra Zero-Shot e Fine-Tuning; tuttavia, non è propriamente un strategia di addattamento dei pesi del modello, perchè essi non vengono modificati. Il modello riceve durante l'inferenza alcuni esempi appartenenti al dataset target, che vengono utilizzati come contesto per produrre la predizione. Sia

$$C = (x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k) \tag{3.11}$$

l'insieme dei k esempi forniti come contesto. La predizione di una nuova istanza x può essere rappresentata come

$$f_{\theta^*}(x \mid C)\,. \tag{3.12}$$

A differenza del fine-tuning, il contesto C non determina un aggiornamento permanente dei parametri θ^*. Gli esempi vengono invece presentati al modello come parte dell'input e consentono di condizionare la predizione sul task specifico [36, 34].

L'ICL può quindi essere interpretato come una forma di adattamento *implicito*: il modello può adattarsi a nuovi task durante la fase di inferenza senza richiedere una nuova ottimizzazione dei pesi.

Una caratteristica fondamentale dell'ICL è rappresentata dalla scelta degli esempi che costituiscono il contesto. Il numero, la qualità e la rappresentatività degli esempi possono influenzare significativamente la qualità delle predizioni. Ad esempio, aumentando il numero di esempi disponibili si può fornire al modello una descrizione più completa del task, ma ciò comporta anche un maggiore costo computazionale e, nei modelli con una

finestra di contesto limitata, il rischio di raggiungere il limite massimo di informazioni elaborabili contemporaneamente.

3. **Fine-tuning**. In questo caso il dataset target viene utilizzato per aggiornare i parametri del modello pre-addestrato. I parametri ottenuti durante il pre-training vengono ulteriormente ottimizzati minimizzando una funzione di perdita definita sul dataset target. Questo permette di sfruttare le rappresentazioni già apprese e, al contempo, di specializzare il modello rispetto alla distribuzione del dataset target.

   Il fine-tuning può essere effettuato secondo differenti livelli di aggiornamento dei parametri. Nel caso del **Full Fine-Tuning**, tutti i parametri del modello vengono aggiornati. Sebbene questa strategia offra la massima libertà di adattamento, il costo computazionale e il consumo di memoria possono diventare rilevanti, soprattutto nel caso di modelli caratterizzati da un numero elevato di parametri. Inoltre, quando il dataset target è di dimensioni ridotte, l'aggiornamento dell'intero modello può aumentare il rischio di overfitting e di perdita di parte delle conoscenze acquisite durante il pre-training.

   Per questo motivo sono stati sviluppati approcci di **Parameter-Efficient Fine-Tuning (PEFT)**, nei quali solamente una piccola parte dei parametri viene ottimizzata mentre il resto del modello rimane congelato [37].

Paradigmi implementati su `TabTune`

La libreria `TabTune` implementa le diverse strategie precedentemente descritte e mette a disposizione una guida che illustra nel dettaglio ciascuna strategia, inclusi i casi in cui utilizzarla e i relativi tradeoffs [31].

Tabella 3.4: Confronto tra le strategie di adattamento di `TabTune`.

Strategia	Training	Caso d'uso	Memoria	Velocità	Accuratezza
Inference	Nessuno	Baseline, zero-shot	Minima	Veloce	Baseline
Fine-Tuning	Tutti i parametri	Elevata accuratezza, risorse ampie	Elevata	Lenta	Massima
PEFT	Adapter LoRA	Memoria limitata, iterazione	Bassa	Media	Elevata

L'utilizzo della strategia può essere configurato attraverso il parametro `tuning_strategy` che può assumere i valori `inference`, `baseft` e `peft`. In particolare:

- **Zero-Shot Inference** (`inference`). In questa modalità vengono utilizzati direttamente i pesi pre-addestrati del modello, senza effettuare alcun aggiornamento dei parametri sul dataset target. Il metodo `fit()` della pipeline viene comunque utilizzato per eseguire le operazioni di preprocessing necessarie, ma non avvia alcun processo di addestramento del modello.
  Questa strategia è indicata per ottenere rapidamente una baseline, valutare le prestazioni del modello pre-addestrato senza adattamento oppure verificare il corretto funzionamento della pipeline di preprocessing.
  Il principale vantaggio consiste nell'assenza di un costo di training, che permette di ottenere predizioni immediatamente e con un consumo di memoria relativamente contenuto. D'altra parte, non essendo previsto alcun aggiornamento dei parametri, il modello non può adattarsi a pattern specifici del dataset target e le prestazioni possono risultare inferiori rispetto alle strategie di fine-tuning. [31]

- **Full Fine-Tuning** (`base-ft`). Questa strategia prevede l'aggiornamento di tutti i parametri del modello utilizzando il dataset target. Il processo di training e l'ottimizzazione dei parametri, viene ripetuto per il numero di epoche specificato, insieme ad una serie di parametri contenuti in `tuning_params`. Tra i principali parametri dispo-

nibili rientrano il dispositivo di esecuzione (`device`), il numero di epoche (`epochs`), il learning rate (`learning_rate`), la dimensione del batch (`batch_size`) e l'ottimizzatore (`optimizer`). [31].

Un esempio di configurazione è il seguente:

```
pipeline = TabularPipeline(
    model_name="TabICL",
    tuning_strategy="base-ft",
    tuning_params={
    "device": "cuda",
    "epochs": 5,
    "learning_rate": 2e-5,
    "batch_size": 32
    }
)


pipeline.fit(X_train, y_train)
predictions = pipeline.predict(X_test)
```

Sebbene il full fine-tuning possa portare a miglioramenti delle prestazioni predittive, richiede un numero maggiore di esempi di addestramento e comporta costi computazionali significativamente più elevati [31].

- **Parameter-Efficient Fine-Tuning (PEFT) mediante LoRA (Low-Rank Adaptation)**

  Gli approcci PEFT consentono di adattare il modello aggiornando solamente una piccola porzione dei parametri, mantenendo congelata la maggior parte della rete pre-addestrata. Questa strategia riduce il numero di parametri ottimizzati e il consumo di memoria, preservando al contempo gran parte della conoscenza acquisita durante il pre-training.

La libreria implementa questa strategia attraverso gli adapter LoRA (*Low-Rank Adaptation*), che introducono matrici addizionali a basso rango da addestrare durante il fine-tuning. Anzichè aggiornare direttamente $\mathbf{W}$, LoRA introduce un aggiornamento:

$$\mathbf{W}' = \mathbf{W} + \Delta\mathbf{W}, \tag{3.13}$$

dove $\mathbf{W}$ rappresenta i pesi originali, mantenuti congelati, mentre $\Delta\mathbf{W}$ rappresenta l'adattamento.

L'aggiornamento $\Delta\mathbf{W}$ viene scomposto nel prodotto di due matrici di dimensioni inferiori:

$$\Delta\mathbf{W} = \mathbf{BA}, \tag{3.14}$$

dove $\mathbf{A}$ e $\mathbf{B}$ sono le matrici a basso rango che vengono ottimizzate durante il fine-tuning. Se $\mathbf{W}$ ha dimensione $d \times k$ e viene scelto un rango r molto inferiore rispetto a d e k, le matrici introdotte da LoRA hanno dimensioni $r \times k$ e $d \times r$. Il numero di parametri da addestrare passa quindi da $d \times k$ a:

$$r(k + d), \tag{3.15}$$

con $r \ll \min(d, k)$.

La configurazione degli adapter viene invece definita attraverso il dizionario `peft_config`, che consente di controllare principalmente il rango delle matrici r. Esso rappresenta un compromesso tra numero di parametri addestrabili e capacità espressiva dell'adattamento. La documentazione indica $r = 8$ come valore predefinito e suggerisce valori inferiori quando si privilegiano compressione e velocità, oppure valori maggiori quando si desidera una maggiore capacità di adattamento. Altri parametri che possono essere

modificati sono: `lora_alpha` che determina il fattore di scala dell'aggiornamento LoRA e `lora_dropout` che introduce una forma di regolarizzazione durante l'addestramento [31].

Un esempio di configurazione è:

```
pipeline = TabularPipeline(
    model_name="TabICL",
    tuning_strategy="peft",
    tuning_params={
        'device': 'cuda',
        'epochs': 5,
        'learning_rate': 2e-4,
        'peft_config': {
            'r': 8,
            'lora_alpha': 16,
            'lora_dropout': 0.05,
        }
    }
)


pipeline.fit(X_train, y_train)
predictions = pipeline.predict(X_test)
```

Durante l'inferenza, il contributo appreso da LoRA viene combinato con i pesi originali del modello [38].

Per i modelli basati su paradigmi di *In-Context Learning* (ICL), la libreria supporta inoltre procedure di *episodic meta-learning*, finalizzate a migliorare la capacità di generalizzazione verso nuovi task e dataset mantenendo le proprietà di adattamento rapido tipiche dei Foundation Models.

Capitolo 4

Design e implementazione della pipeline sperimentale

Questo capitolo descrive come è stata sviluppata la pipeline predittiva finalizzata alla classificazione dei fenotipi fibrotici e non fibrotici nelle neoplasie mieloproliferative, utilizzando inizialmente esclusivamente dati genetici e successivamente integrando variabili cliniche.

L'analisi è stata implementa in Python all'interno dell'ambiente Google Colab, sfruttando librerie dedicate al preprocessing dei dati, all'addestramento dei modelli e alla valutazione delle performance predittive (`NumPy, Pandas, scikit-learn, TabTune`).

Saranno valutati sia differenti approcci di classificazione supervisionata classici, sia i più recenti Tabular Foundation Model.

4.1 Dataset e definizione del task predittivo

Lo studio è stato condotto su un dataset fornito dall'IRCCS Policlinico San Matteo di Pavia che contiene i dati di pazienti affetti da MPN. Il dataset era già stato sottoposto ad un lavoro

preliminare di preprocessing, necessario per poterlo impiegare nello sviluppo di un modello di intelligenza artificiale per la diagnosi di fibrosi nelle neoplasie mieloproliferative.

Il dataset è costituito da 532 pazienti affetti da MPN e comprende un totale di 110 variabili cliniche, laboratoristiche, istopatologiche e genetiche.

Le informazioni raccolte riguardano diverse fasi del percorso clinico del paziente e possono essere raggruppate nelle seguenti categorie:

- **Dati anagrafici e identificativi**: identificativo univoco del paziente (UPN), data di nascita, sesso, professione e altre informazioni demografiche.
- **Dati clinici**: diagnosi iniziale, data di diagnosi, eventi trombotici ed emorragici, eventuali trasformazioni ematologiche, comparsa di neoplasie secondarie, stato in vita e follow-up.
- **Parametri ematologici e biochimici**: valori di leucociti, emoglobina, ematocrito, piastrine, lattato deidrogenasi (LDH), eritropoietina, presenza di splenomegalia e conta delle cellule CD34 circolanti.
- **Caratteristiche molecolari driver**: stato mutazionale dei geni principali associati alle neoplasie mieloproliferative, tra cui *JAK2*, *CALR* e *MPL*.
- **Caratteristiche istopatologiche**: risultati della biopsia osteomidollare, grado di fibrosi midollare e valutazioni della cellularità midollare.
- **Dati genomici avanzati**: risultati del sequenziamento mediante Next Generation Sequencing (NGS), comprendenti numerosi geni frequentemente coinvolti nella patogenesi delle neoplasie mieloproliferative, tra cui *ASXL1*, *DNMT3A*, *TET2*, *SRSF2*, *EZH2*, *RUNX1*, *TP53* e altri.

Le analisi si sono concentrati su pazienti affetti da MPN Ph− quali, trombocitemia essenziale (TE), mielofibrosi prefibrotica (MF0) e mielofibrosi fibrotica (MFI). Nell'analisi non verrà

considerata la policitemia vera (PV) poichè il dataset comprendeva un numero limitato di pazienti sottoposti a sequenziamento NGS.

Il problema è stato formulato come task di classificazione binaria supervisionata. I pazienti sono stati suddivisi in due classi sulla base della diagnosi iniziale:

- **classe fibrotica**: mielofibrosi pre-fibrotica (MFI), mielofibrosi fibrotica (MF0) e Mielofibrosi Familiare
- **classe non fibrotica**: Trombocitemia essenziale (TE) e Trombocitemia essenziale familiare

La variabile target è stata quindi codificata in forma binaria:

- 1 = fenotipo fibrotico
- 0 = fenotipo non fibrotico

4.2 Pre-processing

Prima dell'applicazione dei modelli predittivi, i dati sono stati sottoposti a una fase di preprocessing comprendente la gestione dei valori mancanti, binarizzazione dei dati NGS, la codifica delle variabili categoriche e la normalizzazione delle variabili numeriche, per renderli più adatti ai modelli utilizzati nelle successive analisi.

Non è stato necessario applicare tecniche di bilanciamento delle classi, poiché la distribuzione delle osservazioni tra le classi risultava già adeguatamente equilibrata.

4.2.1 Features genetiche

Il dataset è costituito da informazioni genetiche provenienti sia dai dati di sequenziamento NGS, sia da esami specifici per verificare la presenza del gene driver, indicato sotto forma di feature categorica (JAK2, MLP, CALR o triplo negativo se il paziente non presenta nessuno dei geni driver precedenti). I geni del pannello NGS sono 59 e ogni colonna del dataset

corrisponde a un gene, identificato con nomenclatura `HGNC` (ad es. ASXL1), che riporta il numero di Single Nucletoide Variants (SNV) osservate per paziente. Per ridurre la dispersione dovuta alla variabilità della conta di SNV, questi valori sono stati binarizzati:

$$x_{ij} = \begin{cases} 1 & \text{se il gene } j \text{ presenta almeno una SNV nel paziente } i \\ 0 & \text{altrimenti} \end{cases}$$

Questa scelta consente di capire come il solo effetto della presenza/assenza mutazionale e le co-mutazioni, senza l'introduzione di ipotesi complesse sulla quantità di varianti, influenza l'output.

Inoltre, la matrice binaria paziente $\times$ geni è stata utilizzata come input per algoritmi di proiezione in spazi a bassa dimensionalità, per valutare l'eventuale presenza di gruppi di pazienti caratterizzati da pattern mutazionali simili (sottosezione 5.2.3).

4.2.2 Features cliniche

Nel dataset le features cliniche comprendono informazioni anagrafiche e cliniche, raccolte in diversi momenti del percorso del paziente, sia al momento della diagnosi sia durante le successive fasi di trattamento e follow-up. Ai fini dell'analisi predittiva, sono state selezionate esclusivamente le variabili disponibili al momento della diagnosi, così da utilizzare informazioni antecedenti o contemporanee alla valutazione del fenotipo. Esse comprendono, ad esempio, la data di nascita, il sesso, il valore di leucociti, emoglobina ed altri valori ematologici e istologici (Elenco completo in Appendice D).

Le features selezionate sono prevalentemente di tipo numerico e, in alcuni casi, risultano disponibili solo per una parte dei pazienti.

4.2.3 Gestione dei dati mancanti

L'analisi preliminare dei pattern di missingness ha evidenziato che alcune variabili presentavano valori mancanti non completamente casuali (*Missing Not At Random*, MNAR). In questi casi, l'assenza del dato poteva riflettere decisioni cliniche, come il non sospetto di eventuale evoluzione clinica.

Per gestire i valori mancanti è stata utilizzata una procedura di imputazione basata sulla mediana per le variabili continue e, per ogni variabile, è stato aggiunto un indicatore binario che assume valore 1 quando il dato è mancante e 0 quando è presente. In questo modo, il modello modello preserva l'informazione associata all'assenza del dato, evitando di considerare l'imputazione come una reale osservazione clinica.

Le feature ottenute al termine della fase di pre-processing sono state quindi utilizzate per le successive analisi esplorative, statistiche e predittive.

4.2.4 Forward Stepwise Selection

I Tabular Foundation Models non posseggono una procedura di feature selection integrata, pertanto essa è stata implementata esternamente (vedi Appendice B). La procedura è stata applicata al training set e verrà utilizzata nella fase finale dell'analisi, nell'ambito dell'interpretabilità del modello, al fine di identificare il numero minimo ottimo di feature che contribuiscono maggiormente alle predizioni del modello.

La Forward Stepwise Selection segue un approccio *greedy*, ossia ad ogni iterazione effettua la scelta che massimizza localmente le prestazioni del modello senza riesaminare le decisioni prese nelle iterazioni precedenti. L'algoritmo inizia con un insieme vuoto di feature e, ad ogni passo, considera tutte le variabili non ancora selezionate. Per ciascuna feature candidata viene addestrato il classificatore utilizzando l'insieme delle feature già selezionate più la feature in esame; le prestazioni vengono quindi stimate mediante validazione incrociata (k-fold cross-

validation). La feature che determina il valore più elevato della metrica di valutazione viene aggiunta definitivamente all'insieme delle feature selezionate.

Indicando con F l'insieme completo delle feature disponibili e con S_t il sottoinsieme selezionato all'iterazione t, la feature scelta al passo successivo è definita come

$$f^* = \arg \max_{f \in F \setminus S_t} \text{Score}\left(S_t \cup f\right), \tag{4.1}$$

dove Score$(\cdot)$ rappresenta la metrica utilizzata per valutare il modello; nel presente lavoro è stato impiegato il punteggio F1 medio ottenuto mediante validazione incrociata.

La procedura viene iterata fino a quando l'aggiunta di nuove feature non produce un miglioramento significativo delle prestazioni del modello. Tale criterio consente di individuare un compromesso tra accuratezza predittiva e complessità del modello, evitando l'inclusione di variabili il cui contributo risulta marginale.

Dal punto di vista computazionale, la Forward Stepwise Selection rappresenta un'approssimazione efficiente della ricerca esaustiva (*exhaustive search*). Quest'ultima richiederebbe infatti la valutazione di tutti i possibili sottoinsiemi di feature, pari a

$$2^p - 1, \tag{4.2}$$

dove p indica il numero totale delle variabili disponibili. Tale numero cresce esponenzialmente con il numero di feature, rendendo la ricerca esaustiva rapidamente impraticabile anche per dataset di dimensioni moderate.

Al contrario, la Forward Stepwise Selection richiede la valutazione di un numero di modelli pari a

$$p + (p-1) + \cdots + 1 = \frac{p(p+1)}{2} \tag{4.3}$$

che cresce quadraticamente con il numero di feature. Pur non garantendo il raggiungimento dell'ottimo globale, l'approccio greedy consente di ottenere, nella pratica, un sottoinsieme di variabili caratterizzato da elevate prestazioni predittive con un costo computazionale significativamente inferiore rispetto alla ricerca esaustiva.

4.3 Baseline sperimentale

Il baseline sperimentale è stato definito al fine di quantificare la capacità predittiva dell'informazione genomica (sottosezione 5.3.1) prima dell'integrazione con le informazioni cliniche e dell'impiego di Tabular Foundation Models. Sono stati utilizzati modelli di machine learning tradizionali, quali: Regressione Logistica, Random Forest e CatBoost. Tutti i modelli sono stati addestrati utilizzando lo stesso partizionamento dei dati in training e test set.

La scelta di modelli appartenenti a famiglie differenti consente di confrontare approcci caratterizzati da meccanismi di apprendimento diversi. La Regressione Logistica costituisce un modello relativamente semplice e interpretabile, che assume una relazione lineare tra le feature. È stata pertanto utilizzata come baseline lineare, consentendo di valutare se l'informazione genomica presenti una componente predittiva catturabile senza introdurre interazioni o relazioni non lineari complesse [39].

Random Forest è invece un metodo basato sulla combinazione di numerosi alberi decisionali. Rispetto alla Regressione Logistica, consente di modellare relazioni non lineari e interazioni tra variabili senza richiedere una specifica definizione a priori della loro forma. Tale caratteristica risulta rilevante nel caso dei dati genomici, nei quali l'effetto combinato di più alterazioni può risultare più informativo della singola mutazione considerata isolatamente [40].

Infine, CatBoost [41] appartiene alla famiglia dei metodi di *gradient boosting* e costruisce iterativamente un insieme di alberi per migliorare la capacità predittiva del modello. Inoltre, gestisce in modo efficiente feature categoriche e numeriche, riduce il rischio di overfitting grazie a meccanismi di regolarizzazione e ordered boosting, e ha dimostrato performance competitive in diversi contesti biomedici e genomici.

Tabella 4.1: Configurazione degli iperparametri per i modelli baseline.

Modello	**Parametro**	**Valore**
Regressione Logistica	penalty	L2 (default)
	C	1.0
	random_state	42
Random Forest	n_estimators	300
	max_depth	10
	random_state	42
CatBoost	iterations	200
	depth	3
	learning_rate	0.05
	subsample	0.8
	random_seed	42
	loss_function	Logloss
	eval_metric	Logloss
	verbose	0

4.4 Integrazione clinico-omica mediante modello `TabPFN`

I modelli classici utilizzano esclusivamente informazioni genomiche, rappresentate dalle mutazioni NGS e dai principali geni driver delle neoplasie mieloproliferative. Sebbene tali approcci abbiano mostrato una discreta capacità discriminante, le prestazioni ottenute (sottosezione 5.3.1) suggeriscono che la sola componente molecolare non sia sufficiente a descrivere completamente la complessità biologica associata allo sviluppo della fibrosi.

Per questo motivo, sono state integrate informazioni cliniche raccolte alla diagnosi (sottosezione 4.2.2) e per valutare il loro contributo nella predizione della diagnosi fibrotica, è stato utilizzato il modello TabPFN [33, 8], progettato specificamente per dataset tabellari di dimensioni piccole o medie (tipicamente fino a circa 10k campioni), regime in cui risulta competitivo rispetto ai principali approcci di machine learning tradizionalmente usati, in termini di performance predittiva.

Nel nostro contesto, caratterizzato da un numero limitato di pazienti e da un elevato numero di feature genomiche e cliniche, questa proprietà è particolarmente rilevante, poiché consente

di sfruttare al meglio il segnale disponibile senza ricorrere a modelli ad alta capacità che richiederebbero cohorti più ampie per un addestramento stabile.

Inoltre, l'assenza di training specifico per task rende TabPFN adatto alla valutazione e al confronto delle performance del modello, in diverse configurazioni di features di input e si minimizza il rischio che eventuali differenze siano dovute a scelte di tuning.

Sono state considerate 3 configurazioni di variabili di input: solo dati genomici, solo clinici, genomici e clinici (sottosezione 5.3.2).

Quest'ultima configurazione, è stata applicata ad ulteriori Tabular Foundation Models al fine di confrontarne le prestazioni. Per questi modelli è stato, inoltre, valutato l'effetto dell'integrazione degli embedding `Mut2Vec`, confrontando la rappresentazione genomica binaria con quella basata sugli embedding.

4.5 Integrazione degli embedding mutazionali

Nel nostro dataset, i geni sono identificati tramite nomenclatura `HGNC` (ad es. ASXL1), mentre nel file degli embedding pre-trained rilasciato dagli autori di `Mut2Vec`, i geni sono identificati tramite `Ensembl Gene ID`. Per cui, è stato necessario cambiare convenzione del nostro dataset.

È stato quindi effettuato un processo di conversione dei simboli genici nei corrispondenti identificativi `Ensembl` utilizzando il pacchetto Python `mygene`, che consente di interrogare database genomici aggiornati e ottenere mapping affidabili tra diverse convenzioni di annotazione.

Come già precedentemente illustrato, gli embedding mutazionali patient-level, derivati da embeddings gene-level `Mut2Vec`, consentono di sintetizzare in un insieme compatto di variabili numeriche informazioni complesse relative alle relazioni tra geni mutati, andando oltre la semplice rappresentazione binaria delle mutazioni dei geni NGS. Tali vettori possono essere utilizzati come feature aggiuntive nei modelli, consentendo di incorporare informazioni

implicite sulle associazioni biologiche e sui pattern di co-mutazione appresi durante la fase di addestramento di `Mut2Vec`.

Una scelta metodologica rilevante riguarda l'eventuale inclusione del gene driver all'interno del processo di embedding. Nel presente lavoro si è valutato che l'informazione relativa alla mutazione su uno dei gene driver è ridondante: il gene identificato come driver è compreso sia tra i geni analizzati dal pannello NGS sia come variabile categorica aggiuntiva.

Per tale motivo, la generazione degli embedding è stata effettuata utilizzando i geni del pannello NGS escludendo i geni identificati come driver. In questo modo, la rappresentazione latente appresa dal modello descrive il contesto mutazionale aggiuntivo del paziente, evitando che il segnale associato al gene driver influenzi eccessivamente l'embedding polarizzandolo.

Inoltre, l'esclusione dei geni driver dalla costruzione degli embedding evita la necessità di introdurre schemi di pesatura specifici per bilanciare il contributo del driver rispetto alle altre mutazioni. In assenza di tale accorgimento, il gene driver potrebbe limitare la capacità dell'embedding di catturare informazioni complementari derivanti dalle co-mutazioni.

Quindi, il gene driver verrà considerata come una variabile categorica nelle features di input al modello.

La valutazione dell'impatto degli embedding pre-trained generati da `Mut2Vec` sulle prestazioni predittive dei modelli TFMs, nelle diverse strategie di tuning, sarà utile per capire se le rappresentazioni apprese da `Mut2Vec` siano complementari a quelle apprese nei TFMs e permetterà di capire meglio quando e come integrare embedding pre-trained nelle pipeline predittive. Questo confronto è rilevante perché i TFMs apprendono già rappresentazioni interne delle feature, quindi, non è scontato che vettori esterni aggiungano valore.

Oltre al confronto diretto delle metriche, è stata condotta un'analisi statistica dei risultati al fine di verificare se le differenze osservate tra i modelli fossero statisticamente significative (sezione 5.4).

4.6 Estensione dell'analisi ad altri modelli della libreria TabTune

Grazie alla libreria TabTune, confrontare i diversi TFM è reso più semplice grazie al tool integrato di benchmarking TabularLeaderboard [31].

Nella sezione User Guide della documentazione del framework (Appendice C), è presente una guida per scegliere i modelli da testare; in base alla grandezza del dataset, costi computazionali e alla tipologia delle features, sono stati presi in esame i seguenti modelli TFM:

- **TabPFN**: il modello è pre-addestrato su un ampio insieme di dataset sintetici generati da una distribuzione a priori e apprende a simulare il processo di inferenza bayesiana. Durante l'utilizzo, è in grado di produrre predizioni direttamente a partire dai dati forniti, risultando particolarmente efficace su dataset di piccole e medie dimensioni, come nel nostro caso;

- **TabICL**: applica il paradigma dell'*in-context learning* ai dati tabellari e grazie alla sua architettura gestisce bene grandi dataset e features numeriche e/o miste.

- **Mitra**: è progettato per supportare sia inferenza diretta sia strategie di fine-tuning. Dalle caratteristiche tecniche risulta un modello che riesce a mantenere mantenendo un buon equilibrio tra prestazioni e complessità computazionale ed è adatto per dataset con dimensioni ridotte.

4.6.1 Benchmark delle strategie di tuning

La disponibilità di differenti strategie di adattamento dei pesi consente di valutare il contributo del fine-tuning e di analizzare il comportamento dei modelli nelle diverse strategie, ove applicabili.

Per ciascun modello è stata considerata una configurazione di inferenza diretta (`inference`), utilizzata come riferimento per valutare l'effettivo contributo delle successive strategie di adattamento. In aggiunta alla configurazione di inferenza, sono state valutate, laddove applicabili, strategie di adattamento del modello mediante fine-tuning. Nello specifico, per TabPFN e TabICL è stata considerata una configurazione di full fine-tuning (`base-ft`), mentre per TabICL e Mitra è stata inoltre valutata una strategia di Parameter-Efficient Fine-Tuning (`peft`). La scelta è coerente con il supporto dichiarato da `TabTune` per questi modelli: il framework indica il supporto `peft` di TabPFN come sperimentale e potenzialmente soggetto a problemi di stabilità.

Le configurazioni usate sono riportate nella Tabella 4.2.

Tabella 4.2: Iperparametri utilizzati durante la procedura di cross-validation e benchmarking.

Model	**Strategy**	**Hyperparameters**
TabPFN	inference	Default configuration
TabICL	inference	Default configuration
Mitra	inference	Default configuration
TabICL	PEFT	epochs = 3, LoRA rank (r) = 8, $\alpha = 16$
Mitra	PEFT	epochs = 3, LoRA rank (r) = 8, $\alpha = 16$
TabICL	base-ft	epochs = 3, learning rate = 2×10^{-5}
TabPFN	base-ft	epochs = 3, learning rate = 2×10^{-5}

Per il benchmarking è stata utilizzata la stessa suddivisione dei dati in training e test set per tutti gli esperimenti, garantendo che le differenze osservate nelle prestazioni siano attribuibili ai modelli e alle strategie di adattamento piuttosto che a differenti campionamenti dei dati. Inoltre, è stato fissato il random seed per assicurare la riproducibilità dei risultati.

Poiché il dataset analizzato presenta una numerosità limitata, una singola suddivisione tra training e test set potrebbe produrre stime delle prestazioni soggette a elevata variabilità. In problemi di classificazione binaria, anche un numero ridotto di esempi classificati correttamente o erroneamente può determinare variazioni significative nelle metriche di performance. Di conseguenza, è stata eseguita la 10-fold cross validation per ottenere una valutazione più robusta delle prestazioni dei modelli (Tabella 5.18).

4.7 Ottimizzazione degli iperparametri

Sebbene nella strategia di tuning in modalità *inference* non esista una procedura di ottimizzazione dei parametri, essa risulta importante da valutare per le altre strategie di adattamento.

Poiché il modello TabPFN fine-tuned (`base-ft`) non ha mostrato un miglioramento rispetto alla strategia di `inference` (Tabella 5.17), è stata effettuata un'ottimizzazione degli iperparametri del processo di fine-tuning, con l'obiettivo di individuare la configurazione in grado di massimizzare le prestazioni del modello (sezione 5.5).

Sono stati valutati gli iperparametri riportati di seguito:

- **n_estimators**: numero di modelli utilizzati nell'ensemble durante l'inferenza. L'impiego di un numero maggiore di estimatori può migliorare la robustezza e la stabilità delle predizioni, a fronte di un incremento del costo computazionale. Sono stati considerati i valori: {8, 16, 32}.

- **softmax_temperature**: parametro che regola la distribuzione di probabilità prodotta dal classificatore. Valori inferiori a 1 rendono le predizioni più confidenti, mentre valori superiori producono distribuzioni più uniformi. Sono stati valutati i valori: {0.5, 0.9, 1.5}.

- **epochs**: numero di epoche di addestramento durante la fase di fine-tuning. Un numero maggiore di epoche consente al modello di adattarsi meglio ai dati di training, ma può aumentare il rischio di overfitting. Sono stati testati i valori: {1, 3, 5}.

- **learning_rate**: tasso di apprendimento utilizzato per l'aggiornamento dei pesi durante il fine-tuning. Questo parametro controlla l'entità delle modifiche apportate ai pesi del modello a ogni iterazione di ottimizzazione. Sono stati considerati i valori: $\{1 \times 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}\}$.

Come suggerito nella documentazione di `TabTune` per trovare la miglior configurazione dei parametri che massimizza la performance del modello nel task specifico, si possono utilizzare 3 metodi diversi: grid search, random search e Bayesian optimization con Optuna.

La selezione della configurazione ottimale è stata effettuata confrontando le prestazioni ottenute dalle diverse combinazioni di iperparametri, con strategia di **Grid Search** su validation set.

4.8 Metriche di valutazione e validazione dei modelli

Per garantire la coerenza metodologica e la comparabilità dei risultati, le prestazioni dei modelli sono state valutate utilizzando le stesse metriche descritte nei capitoli precedenti. In particolare, per il task di classificazione binaria sono state considerate le metriche di Accuracy, Precision, Recall, F1-score e AUC-ROC.

Tutte le metriche sono state ottenute grazie alla libreria Python `Scikit-Learn`.

Infine, anche in questo caso per garantire una valutazione robusta e affidabile delle prestazioni dei modelli, è stata adottata una procedura di 10-fold cross-validation.

4.9 Analisi di interpretabilità

Sebbene il modello con le migliori prestazioni complessive sia risultato `TabPFN` addestrato sulle rappresentazioni ottenute tramite embedding `Mut2Vec`, l'analisi di interpretabilità e spiegazione del modello, è stata eseguita sul modello `TabPFN` sulle feature originali. Questa scelta è motivata dal fatto che gli embedding trasformano le mutazioni in una rappresentazione latente, le cui componenti non sono direttamente associabili alle caratteristiche biologiche originali, rendendo l'interpretazione dell'importanza delle singole mutazioni poco significativa.

I risultati dovranno, quindi, essere interpretati come una misura dell'importanza delle feature originali per `TabPFN` in assenza della trasformazione `Mut2Vec` e non come una spiegazione

diretta del modello finale basato sugli embedding.

4.9.1 Permutation Feature Importance

L'analisi di interpretabilità è stata condotta mediante Permutation Feature Importance. Questo metodo costituisce una tecnica model-agnostic, applicabile anche ai Tabular Foundation Models indipendentemente dalla loro specifica architettura. La tecnica valuta il contributo predittivo di una variabile attraverso la variazione delle prestazioni del modello conseguente alla permutazione dei suoi valori [42]. Tuttavia, è importante considerare che, nel caso dei Tabular Foundation Models, il costo computazionale della PFI può risultare maggiore rispetto ai modelli tradizionali, poiché la procedura richiede numerose valutazioni del modello durante l'inferenza [43].

Per ridurre la variabilità dovuta alla casualità della permutazione, l'operazione è stata ripetuta 10 volte e i valori riportati (sezione 5.6) rappresentano la media delle importanze ottenute.

In una fase preliminare, è stata eseguita la PFI sull'intero set di feature: questa prima analisi aveva uno scopo esplorativo e che ha evidenziato un'importanza elevata non solo per variabili clinicamente interpretabili, ma anche per alcune feature legate alla gestione dei dati mancanti e alla loro imputazione (Tabella 5.25e Tabella 5.28). Non essendo stata ottenuta un'interpretazione sufficientemente selettiva del comportamento del modello, è stata eseguita la feature selection.

Per individuare il sottoinsieme ottimale di variabili è stata adottata una procedura di Forward Stepwise Feature Selection, una tecnica di selezione incrementale delle feature sulla base delle prestazioni del classificatore, addestrandolo ripetutamente su differenti sottoinsiemi di variabili (sottosezione 4.2.4). I risultati sono riportati in Tabella 5.26.

Successivamente è stata rieseguita la PFI sul modello addestrato sul numero ottimo minimo di feature selezionate (Tabella 5.27).

Infine, la PFI sul modello addestrato su dati clinici e genomici, ha evidenziato che le variabili cliniche costituiscono i principali fattori discriminanti, relegando il contributo delle alterazioni genomiche a un ruolo secondario. Tale comportamento è atteso, poiché i parametri clinici descrivono direttamente il fenotipo della malattia e risultano quindi maggiormente correlati alla classificazione finale. Per valutare il reale contributo delle informazioni molecolari, è stata condotta un'ulteriore analisi limitata alla componente genomica (Tabella 5.31).

4.10 Valutazione del contributo della Biopsia Ossea

L'analisi della distribuzione del grado di fibrosi midollare ha evidenziato una marcata associazione tra i gradi più elevati di fibrosi e la diagnosi di mielofibrosi. I pazienti con grado di fibrosi pari a 2 o 3 risultavano esclusivamente affetti da mielofibrosi, mentre i pazienti con grado 0 o 1 comprendevano sia casi di trombocitemia essenziale sia casi di mielofibrosi iniziale.

Tabella 4.3: Distribuzione della diagnosi iniziale in funzione del grado di fibrosi midollare.

BOMFibrosi	**TE**	**MF0**	**MFI**	**Totale**
0	221	40	1	262
1	14	52	46	112
2	0	1	59	60
3	0	1	39	40
Missing	10	9	36	55
Totale	245	103	181	529

Questa distribuzione (Tabella 4.3) suggerisce quindi, che la presenza di un grado di fibrosi midollare elevato, risultante dalla biopsia osteo-midollare, rappresenta un forte indicatore diagnostico e potrebbe determinare una sovrastima delle prestazioni del modello predittivo, in quanto i casi con fibrosi severa risultano relativamente semplici da classificare. Pertanto, la distinzione diagnostica risulta più complessa nelle fasi iniziali della malattia, rendendo necessaria un'analisi limitata ai pazienti con grado di fibrosi pari a 0 o 1, così da verificare la robustezza del modello in scenari clinicamente eterogenei.

Specificamente, sono state condotte le seguenti analisi:

- confronto delle performance dei modelli `TabTune` considerati, escludendo le variabili direttamente derivate dalla biopsia ossea (Tabella 5.22), al fine di valutare la capacità discriminativa delle sole informazioni cliniche e laboratoristiche disponibili prima dell'esecuzione della biopsia;

- analisi di sensibilità (sezione 5.7), verificando la capacità del miglior modello osservato, di discriminare la diagnosi di fibrosi, dapprima limitatamente ai soli pazienti con grado di fibrosi midollare pari a 0 o 1, escludendo i pazienti con fibrosi avanzata (gradi 2 e 3), e successivamente considerando l'intera coorte;

4.10.1 Caratterizzazione dei sottotipi fenotipici

La caratterizzazione dei sottogruppi ha lo scopo di mostrare che il modello è clinicamente significativo e che è in grado di stratificare i pazienti in gruppi di rischio coerenti, senza considerare le informazioni cliniche ottenute dalla biopsia, poichè esse sono strettamente correlate all'outcome e potrebbero determinare una sovrastima delle prestazioni del modello. Inoltre, si vuole valutare quanto il modello possa essere utile in una fase precoce della pratica clinica.

Per ciascuno dei pazienti appartenenti al test set è stata calcolata la probabilità individuale dell'outcome utilizzando il modello predittivo ottimo. Le probabilità ottenute sono state successivamente ordinate in senso crescente e suddivise in quattro gruppi di uguale numerosità (quartili), corrispondenti a livelli progressivamente crescenti di rischio predetto (sezione 5.8).

La suddivisione in quartili rappresenta un approccio comunemente impiegato nella valutazione dei modelli predittivi, in quanto consente di classificare la popolazione in gruppi omogenei sulla base del rischio stimato e di verificare se il modello sia in grado di identificare sottopopolazioni con differenti caratteristiche cliniche e differente prevalenza dell'outcome [44].

La stratificazione ha permesso di valutare non soltanto la prevalenza osservata della fibrosi nei diversi livelli di rischio, ma anche la distribuzione delle principali caratteristiche cliniche (Tabella 5.35) dei pazienti: sono state considerate età alla diagnosi, sesso, dimensione splenica, emoglobina, leucociti, piastrine, LDH e CD34 circolanti. La selezione di tali variabili è stata motivata dalla loro rilevanza clinica nel contesto delle neoplasie mieloproliferative e dalla loro possibile associazione con la presenza e la progressione della fibrosi.

Una prima analisi basata sulle mediane delle variabili quantitative aveva tuttavia evidenziato alcune criticità interpretative, soprattutto per le variabili caratterizzate da un'elevata percentuale di valori mancanti e successivamente imputati. In tali condizioni, la mediana può risultare poco rappresentativa della distribuzione effettiva dei valori osservati: ad esempio, nel caso dei CD34 circolanti, l'elevata presenza di valori imputati alla mediana globale determinava valori mediani identici nei diversi quartili di rischio, pur in presenza di distribuzioni differenti.

Per questo motivo, le variabili quantitative sono state discretizzate mediante soglie clinicamente interpretabili. Per ciascuna categoria sono stati quindi calcolati il numero di pazienti e la relativa percentuale all'interno di ciascun quartile di rischio.

L'età alla diagnosi è stata mantenuta come variabile quantitativa e descritta mediante la mediana nei quattro quartili, il sesso è stato invece considerato come variabile categorica e descritto mediante frequenze assolute e percentuali.

Il p-value è ottenuto mediante test di Kruskal–Wallis per l'età e mediante test χ^2 per le variabili categorizzate.

Infine, un'attenzione particolare è stata dedicata ai pazienti con probabilità predetta compresa tra 0,40 e 0,60, che rappresentano una "zona grigia" decisionale in cui il modello mostra minore confidenza. Analizzare questi casi borderline consente di capire se gli errori di classificazione si concentrino in questa fascia e se questi pazienti presentino profili clinici o demografici particolari, fornendo indicazioni utili sull'affidabilità del modello e sulle situazioni in cui la sua applicazione richiede maggiore cautela (Tabella 5.36).

Capitolo 5

Risultati

In questo capitolo verranno presentati i risultati ottenuti nel corso dell'analisi sperimentale.

In una prima parte sarà descritta la popolazione oggetto di studio e i principali risultati relativi alla caratterizzazione genomica dei pazienti, attraverso l'analisi della distribuzione delle alterazioni, delle associazioni tra geni e della loro rappresentazione nello spazio genomico.

Successivamente saranno riportati i valori di performance dei modelli predittivi, dalle diverse configurazioni delle informazioni cliniche e genomiche, al confronto tra i Tabular Foundation Models considerati e le strategie di fine-tuning.

Saranno quindi approfonditi i risultati dell'analisi di interpretabilità mediante Permutation Feature Importance. Infine, sono presentate le analisi volte a valutare il contributo della biopsia osteomidollare e la capacità delle probabilità predette dal modello di stratificare i pazienti in sottogruppi fenotipici ben identificabili.

5.1 Caratteristiche della popolazione in studio

La popolazione complessivamente analizzata comprende 532 pazienti affetti da neoplasie mieloproliferative. La diagnosi iniziale più rappresentata è la TE, presente in 245 pazienti

(46,1%). Segue la MFI, diagnosticata in 181 pazienti (34,0%), mentre la MF0 è presente in 103 pazienti (19,4%). Le forme familiari erano rappresentate da 2 casi di TE familiare (0,4%) e da 1 caso di mielofibrosi familiare (0,2%).

La distribuzione delle principali caratteristiche demografiche della coorte è riportata nella Tabella 5.1.

Tabella 5.1: Principali caratteristiche demografiche della popolazione studiata.

Caratteristica	**Valore**
Numero totale di pazienti	532
Età alla diagnosi	
Media ± DS	55,0 ± 15,7
Mediana (IQR)	56,9 (43,8–67,2)
Sesso,(M)	254 (47,7%)
Sesso(F)	278 (52,3%)

Per quanto riguarda i principali parametri ematologici disponibili alla valutazione: la conta leucocitaria è disponibile per 528 pazienti (99,2% della coorte), con una mediana di 8,6. L'emoglobina presenta una mediana di 13,2, mentre la conta piastrinica mostra una mediana di 619,5. L'ematocrito è disponibile per 341 pazienti, con una mediana di 40,6. I valori di LDH, disponibili per 415 pazienti, presentano una mediana di 255,9

Tabella 5.2: Principali parametri clinico-laboratoristici della popolazione.

Parametro	**N. pazienti**	**Mediana**	**Intervallo interquartile**	**Range**
Leucociti	528	8,6	7,0–11,3	1,6–36,0
Emoglobina	528	13,2	11,7–14,6	3,4–19,6
Ematocrito	341	40,6	36,7–44,2	10,4–61,0
Piastrine	528	619,5	426,0–810,0	22,0–3279,0
LDH	415	255,9	190,6–372,8	76,7–2410,0
Milza	514	0,0	0,0–2,0	0,0–25,0
CD34 circolanti	289	5,7	3,4–12,3	0,4–1012,5
Eritropoietina	267	8,2	4,5–14,4	0,0–1663,0

Nota: il numero di pazienti varia in funzione della disponibilità del dato per ciascun parametro.

La distribuzione delle principali mutazioni driver nella popolazione studiata è riportata nella Tabella 5.3.

Tabella 5.3: Distribuzione delle principali mutazioni driver nella popolazione studiata.

Mutazione driver	**N. pazienti**
JAK2	310
CALR	139
MPL	27
Triplo negativo	56

L'analisi molecolare mediante NGS è disponibile per tutti i pazienti. Di cui, 252 pazienti (47,3%) non presentano mutazioni aggiuntive rilevate dal pannello di geni considerato, mentre 143 (26,8%) presentavano una mutazione, 80 (15%) due mutazioni, 31 (5,8%) tre mutazioni e 26 (4,9%) quattro più mutazioni.

Tabella 5.4: Numero di mutazioni rilevate mediante NGS.

N. mutazioni	**N. pazienti**
0	252
1	143
2	80
3	31
4+	26

5.2 Caratterizzazione genomica della coorte

L'analisi esplorativa è stata inizialmente condotta con l'obiettivo di caratterizzare il profilo genomico della coorte e di identificare le principali caratteristiche della distribuzione delle alterazioni. Sono state analizzate la frequenza delle alterazioni a livello genico, le associazioni tra geni e la struttura dello spazio genomico mediante tecniche di riduzione dimensionale.

5.2.1 Distribuzione delle alterazioni genomiche

Nel Grafico 5.1 si osservano differenze nel profilo mutazionale tra i gruppi, con una maggiore prevalenza di mutazioni aggiuntive nei pazienti con mutazione sul gene driver MPL e una quota più elevata di casi senza mutazioni nei pazienti triplo negativi (ovvero che non presentano nessun gene driver mutato).

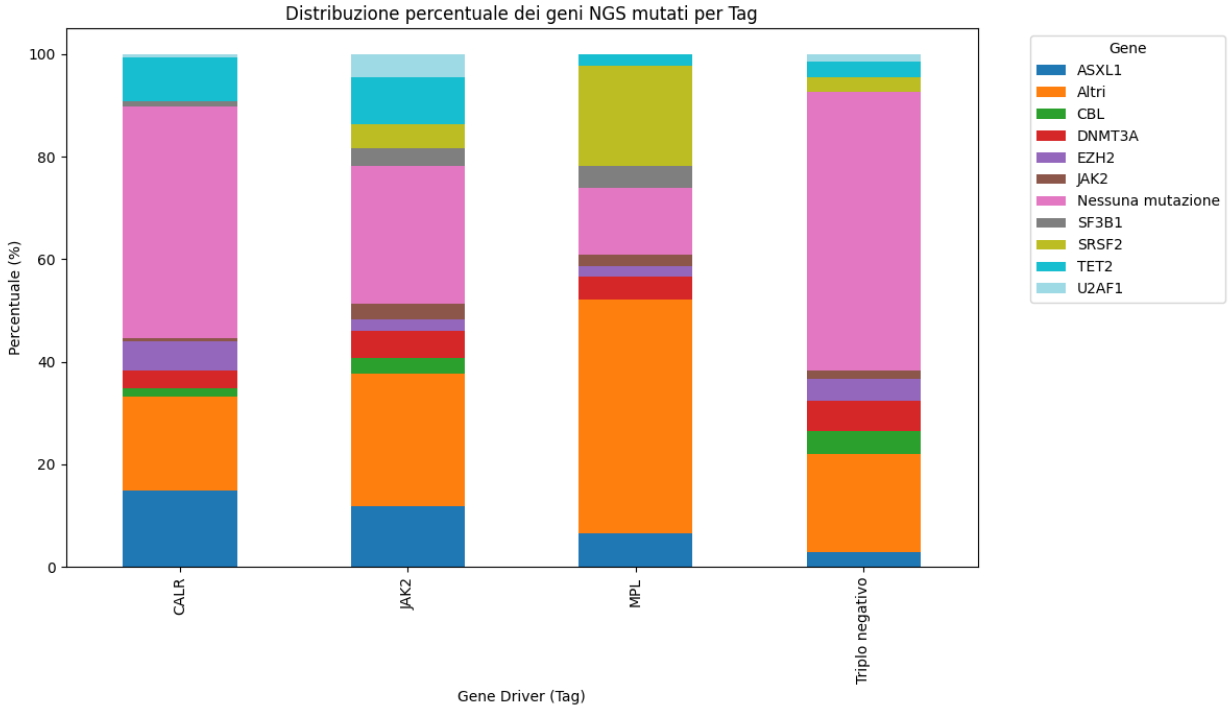

Figura 5.1: Distribuzione percentuale delle mutazioni NGS nei diversi sottogruppi definiti dal gene driver (CALR, JAK2, MPL e triplo negativo). Ogni barra rappresenta la composizione percentuale delle mutazioni osservate per ciascun gruppo, includendo i principali geni mieloidi analizzati (ASXL1, CBL, DNMT3A, EZH2, JAK2, SF3B1, SRSF2, TET2 e U2AF1) e la quota di pazienti senza mutazioni aggiuntive.

Nel Grafico 5.2 è mostrata la distribuzione percentuale delle mutazioni dei geni del pannello NGS nei diversi sottogruppi clinici dei pazienti JAK2-mutati. Le barre mostrano la proporzione relativa dei geni mutati e dei casi senza mutazioni aggiuntive. I profili mutazionali risultano più eterogenei nei sottogruppi MF0 e MFI (mielofibrosi pre-fibrotica e fibrotica) rispetto alla TE/TE familiare, caratterizzata dall'assenza di mutazioni concomitanti.

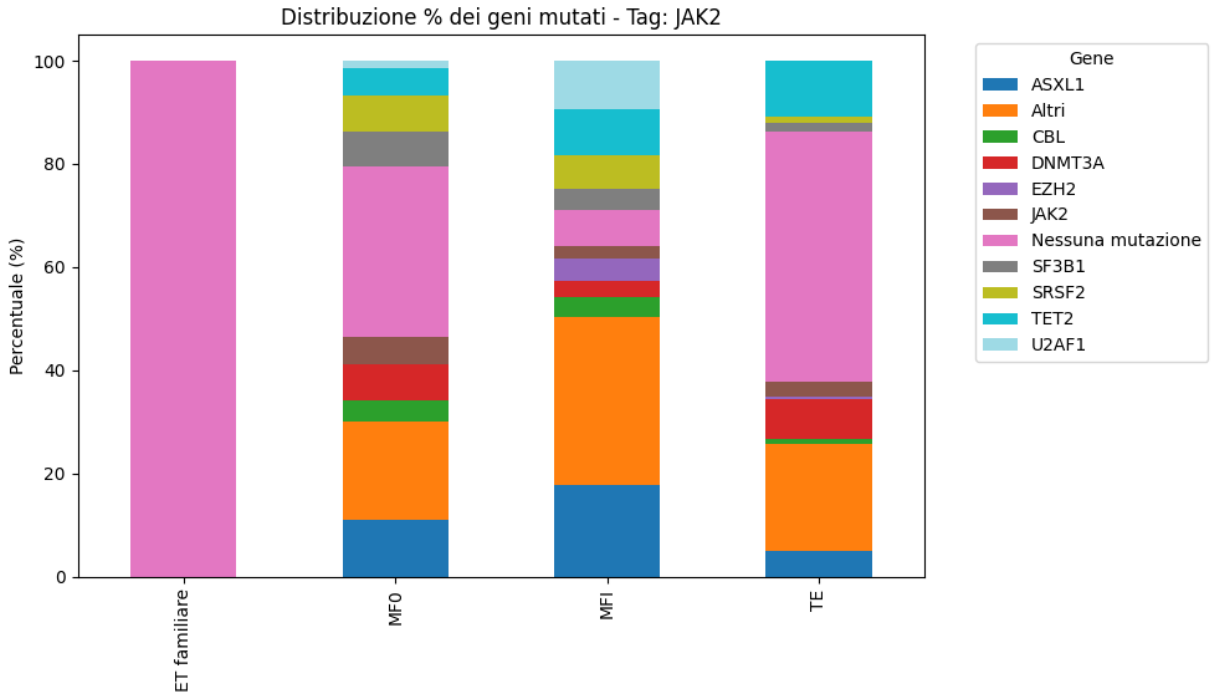

Figura 5.2: Distribuzione percentuale delle mutazioni NGS nei diversi sottogruppi clinici dei pazienti JAK2-mutati.

Tabella 5.5: Principali mutazioni NGS nei pazienti con mutazione driver JAK2.

Gene	Conteggio	Percentuale (%)
Nessuna mutazione	130	26.97
ASXL1	57	11.83
TET2	44	9.13
DNMT3A	26	5.39
U2AF1	22	4.56

Nel caso di pazienti CALR-mutati, la prevalenza di casi senza mutazioni aggiuntive è particolarmente elevata nella mielofibrosi familiare e nel gruppo TE, mentre i pazienti MFI mostrano il profilo mutazionale più eterogeneo, con una maggiore rappresentazione di mutazioni in ASXL1, EZH2, KMT2A e TET2. Questi dati suggeriscono una complessità molecolare variabile tra le diverse manifestazioni cliniche associate alla mutazione driver CALR.

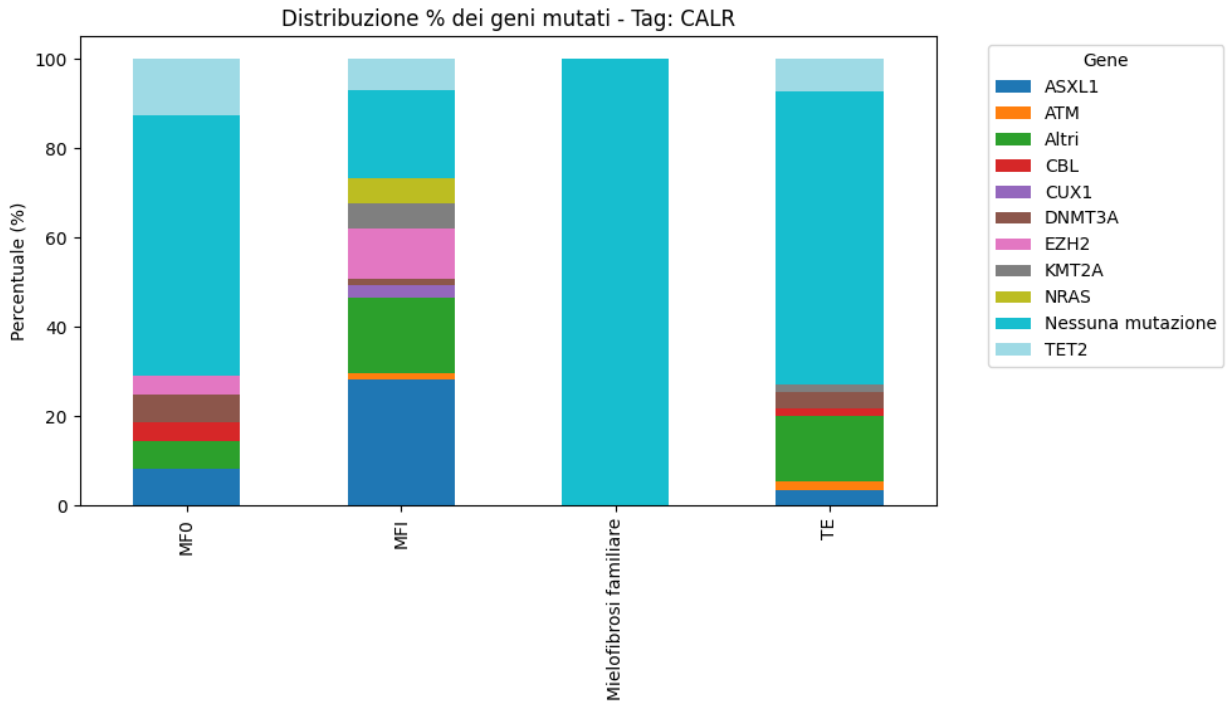

Figura 5.3: Distribuzione percentuale dei geni mutati nei sottogruppi clinici dei pazienti CALR-mutati.

Tabella 5.6: Principali mutazioni NGS nei pazienti con mutazione driver CALR.

Gene	Conteggio	Percentuale (%)
Nessuna mutazione	79	45.14
ASXL1	26	14.86
TET2	15	8.57
EZH2	10	5.71
DNMT3A	6	3.43

Nei pazienti con nessuna mutazione su geni driver rilevata, la quota di casi senza mutazioni aggiuntive è predominante nei gruppi MF0 e TE, mentre il gruppo MFI presenta il profilo mutazionale più eterogeneo.

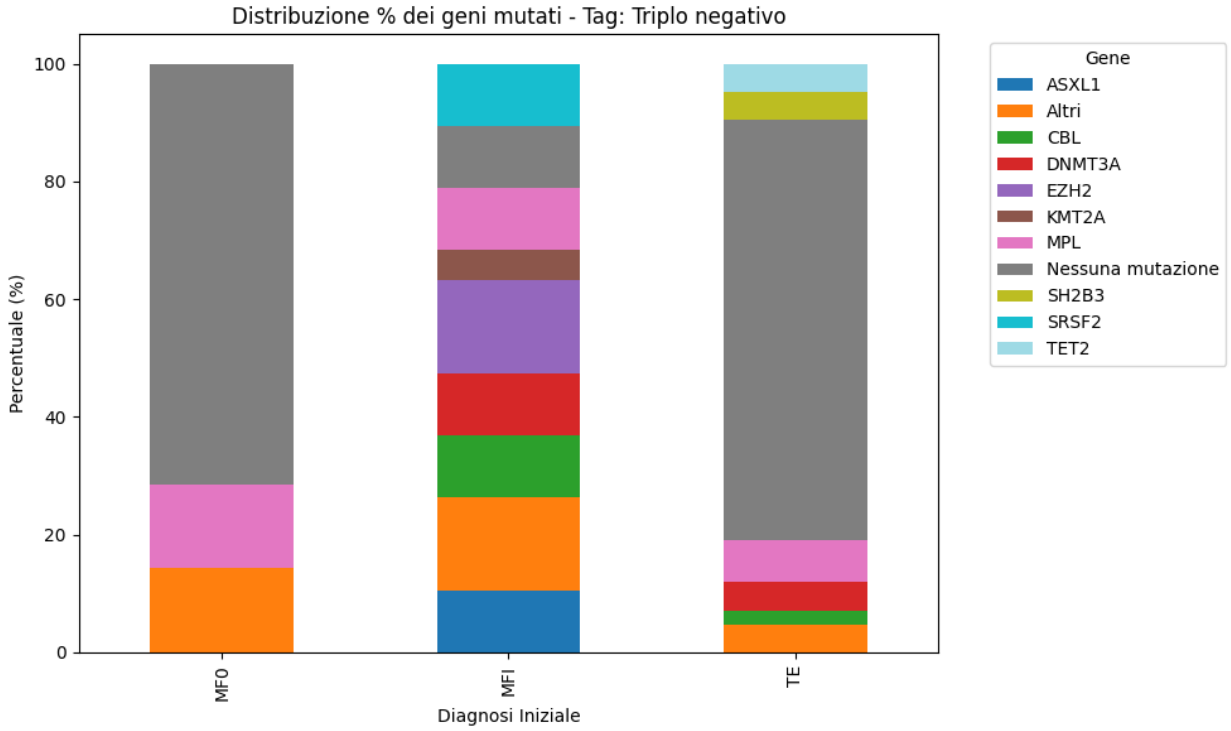

Figura 5.4: Distribuzione percentuale dei geni mutati nei sottogruppi clinici dei pazienti triplo negativi.

Tabella 5.7: Principali mutazioni NGS nei pazienti triplo negativi.

Gene	Conteggio	Percentuale (%)
Nessuna mutazione	37	54.41
MPL	6	8.82
DNMT3A	4	5.88
EZH2	3	4.41
CBL	3	4.41

Infine, nei pazienti MPL-mutati, la TE familiare presenta esclusivamente mutazioni in DNMT3A, mentre i sottogruppi MF0, MFI e TE mostrano una maggiore eterogeneità molecolare. Il gruppo MFI evidenzia la più ampia varietà di mutazioni concomitanti, mentre nel gruppo TE prevalgono i casi senza ulteriori alterazioni genetiche.

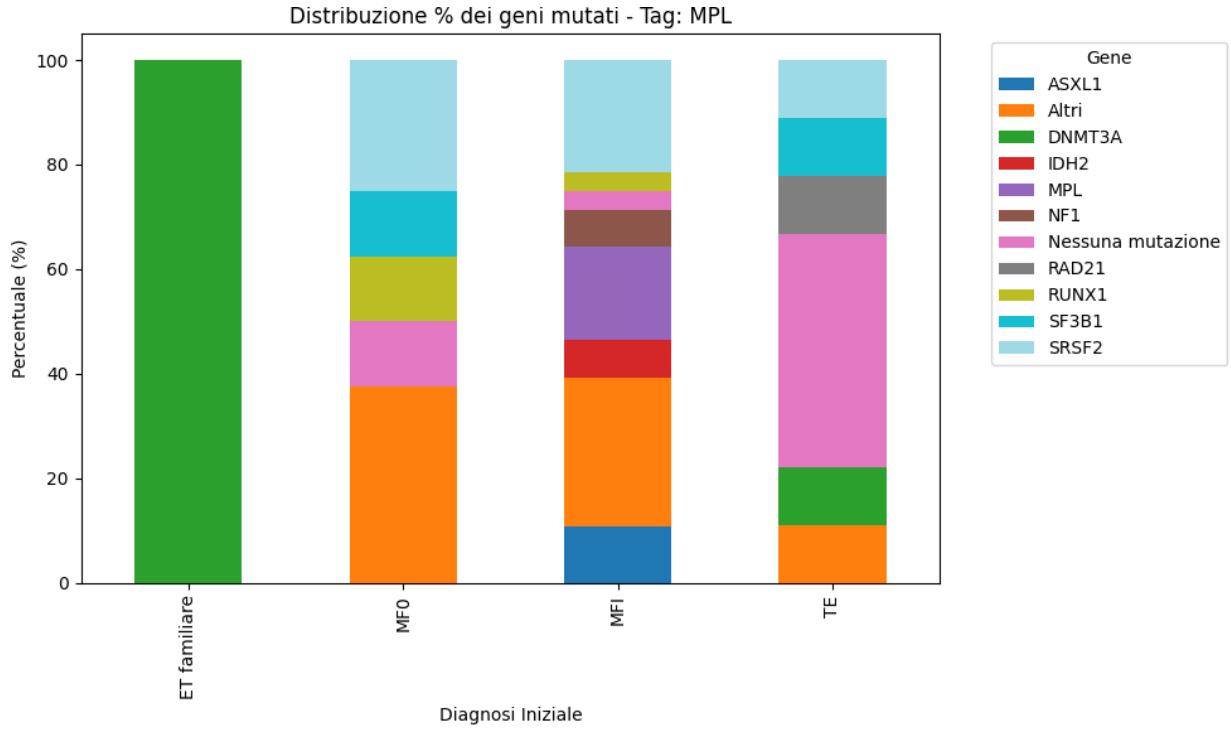

Figura 5.5: Distribuzione percentuale dei geni mutati nei sottogruppi clinici dei pazienti MPL-mutati.

Tabella 5.8: Principali mutazioni NGS nei pazienti con mutazione driver MPL.

Gene	Conteggio	Percentuale (%)
SRSF2	9	19.57
Nessuna mutazione	6	13.04
MPL	5	10.87
ASXL1	3	6.52
NF1	2	4.35

5.2.2 Co-occorrenza e associazione tra geni

I risultati mostrano le principali associazioni mutazionali tra geni sequenziati tramite NGS, esclusi i geni driver e mutazionalmente ultra-rari, fornendo informazioni sui possibili meccanismi biologici coinvolti nella patogenesi e nell'evoluzione clonale della malattia.

Come riportato nella Tabella 5.9, l'associazione più forte è stata osservata tra *NRAS* e *SETBP1*. Tra le altre associazioni rilevanti si osservano quella tra *SRSF2* e *IDH2* e quelle di *ASXL1* con *U2AF1*.

Nel complesso, tutte le associazioni riportate presentano valori di odds ratio superiori a 1, indicando una maggiore frequenza di co-occorrenza delle alterazioni rispetto a quanto atteso in condizioni di indipendenza. Si osserva che le associazioni che coinvolgono il gene *ASXL1* mostrano valori di FDR dell'ordine di 10^{-6}, risultando tra quelle statisticamente più significative dopo la correzione per i confronti multipli.

Tabella 5.9: Principali associazioni mutazionali significative identificate mediante test esatto di Fisher e correzione per confronti multipli secondo Benjamini-Hochberg (FDR).

Gene 1	Gene 2	Odds Ratio	p-value	FDR
ASXL1	EZH2	10.02	2.40×10^{-9}	1.0×10^{-6}
ASXL1	U2AF1	15.36	1.36×10^{-9}	1.0×10^{-6}
ASXL1	CBL	9.75	1.28×10^{-6}	4.23×10^{-4}
NRAS	SETBP1	130.50	2.76×10^{-5}	6.83×10^{-3}
SRSF2	IDH2	68.69	5.93×10^{-5}	9.78×10^{-3}

La heatmap con clustering gerarchico (Figura 5.6) ha evidenziato la presenza di cluster comprendenti geni coinvolti nei processi di splicing dell'RNA (SRSF2, SF3B1, U2AF1 e ZRSR2) e nella regolazione epigenetica (ASXL1, TET2, EZH2 e DNMT3A), suggerendo una tendenza alla co-occorrenza di alterazioni appartenenti alle medesime vie biologiche.

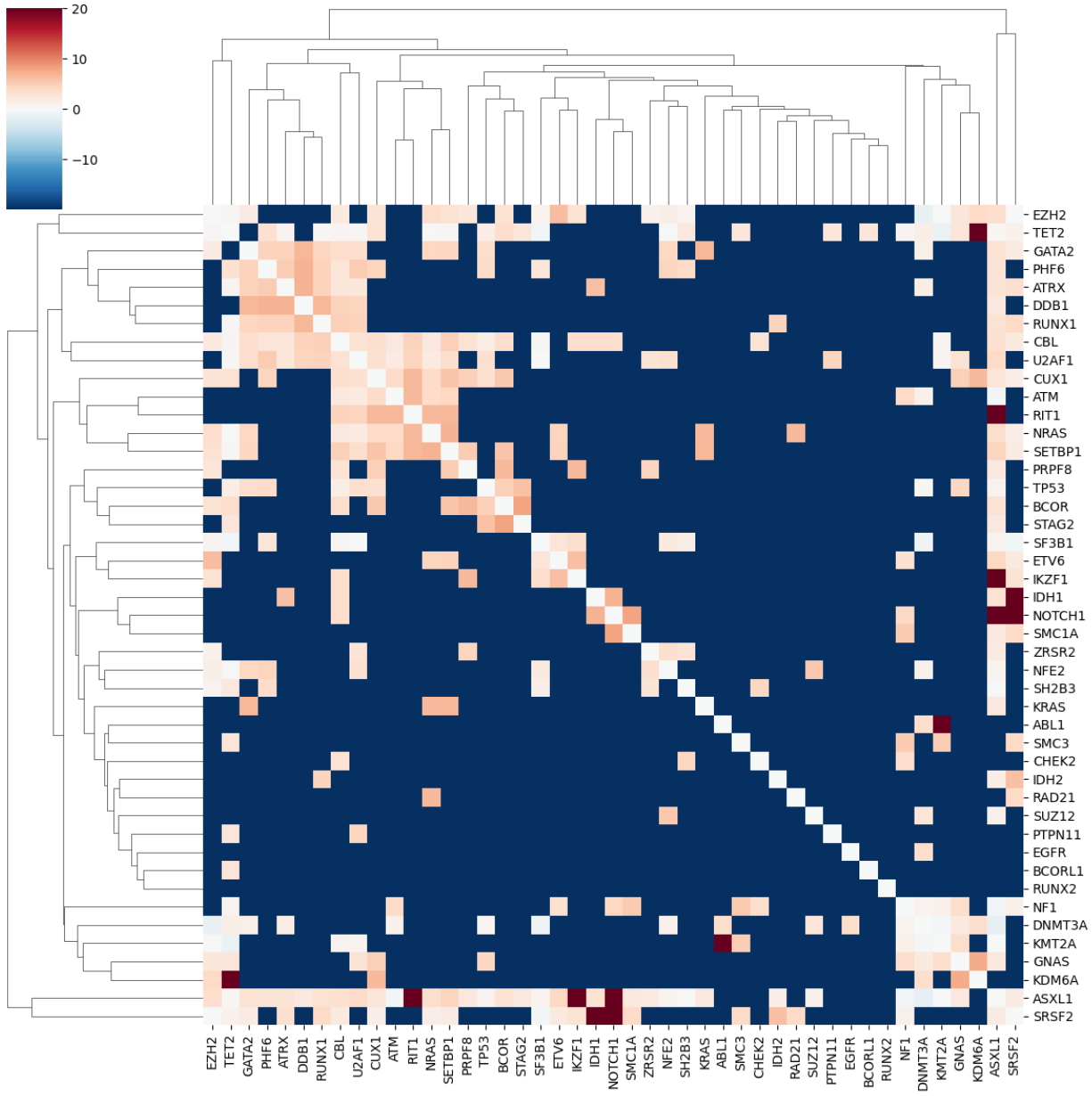

Figura 5.6: Heatmap di co-occorrenza mutazionale ottenuta mediante analisi delle associazioni tra tutte le coppie di geni valutati con sequenziamento NGS. Per ciascuna coppia è stato calcolato l'odds ratio (OR) mediante test esatto di Fisher su tabella di contingenza 2×2; i valori rappresentati nella matrice corrispondono al $\log_2(\mathrm{OR})$. Le tonalità rosse indicano una tendenza alla co-occorrenza delle mutazioni ($\mathrm{OR} > 1$), mentre le tonalità blu indicano una tendenza alla mutua esclusività ($\mathrm{OR} < 1$). L'intensità del colore riflette la forza dell'associazione tra le due alterazioni genetiche.

L'analisi delle associazioni tra mutazioni NGS e categoria del gene driver ha evidenziato pattern distinti. Si osservano alcune associazioni statisticamente significative nella regione relativa ai geni driver MPL/CALR.

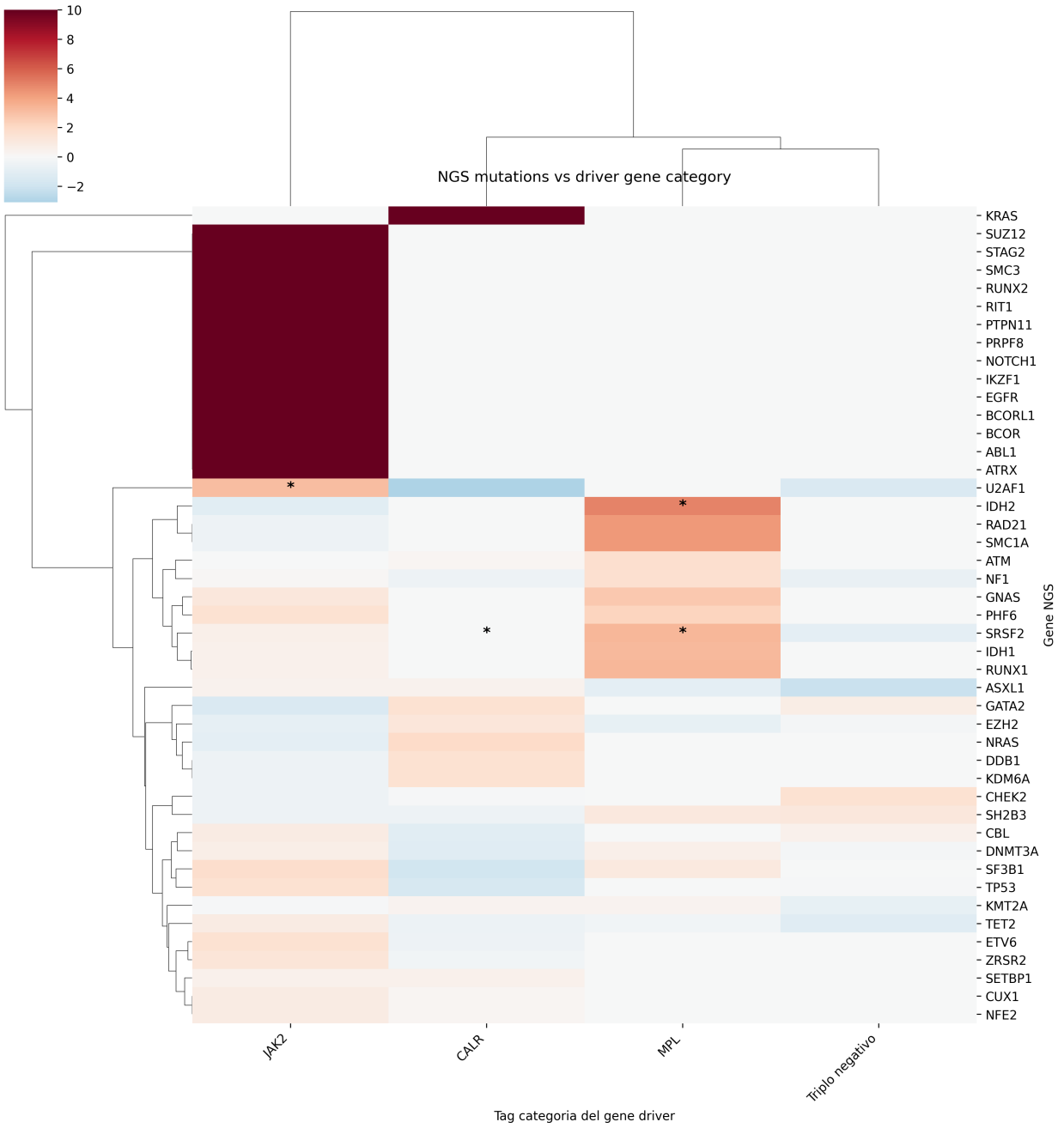

Figura 5.7: Heatmap di co-occorrenza mutazionale ottenuta mediante analisi delle associazioni tra i geni NGS e geni driver. Gli asterischi indicano le associazioni statisticamente significative.

5.2.3 Visualizzazione dello spazio genomico

Come mostrato in Figura 5.8, i pazienti portatori di differenti geni driver (JAK2, CALR e MPL) risultano ampiamente sovrapposti nello spazio delle prime due componenti principali, senza evidenza di cluster chiaramente separati. In particolare:

- i pazienti JAK2-mutati (blu) sono distribuiti lungo tutto il grafico;
- i pazienti CALR-mutati (arancione) non formano un cluster separato;
- i casi MPL (rosso) risultano dispersi e sovrapposti agli altri gruppi;
- i pazienti triplo negativi (verde) occupano regioni condivise con gli altri sottotipi.

Questa distribuzione suggerisce che i profili mutazionali non siano fortemente determinati dal gene driver e che non esistano pattern mutazionali chiaramente distinti associati ai diversi sottotipi molecolari. Sono inoltre osservabili alcuni pazienti isolati rispetto al nucleo principale della distribuzione, verosimilmente caratterizzati da combinazioni mutazionali rari.

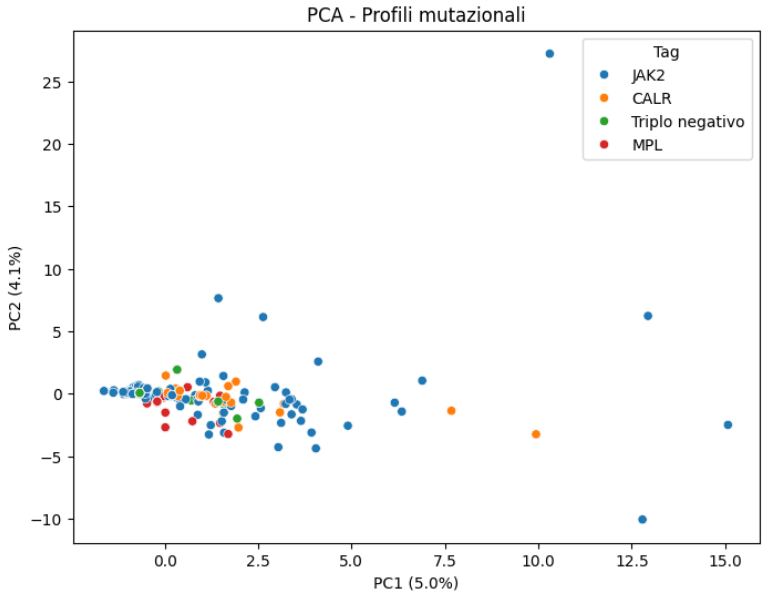

Figura 5.8: Analisi delle componenti principali (PCA) dei profili mutazionali NGS ottenuta escludendo i geni driver principali, colorati per gene driver. Ogni punto rappresenta un paziente ed è colorato in base al gene driver. Le prime due componenti principali spiegano rispettivamente il 5.0% e il 4.1% della varianza totale.

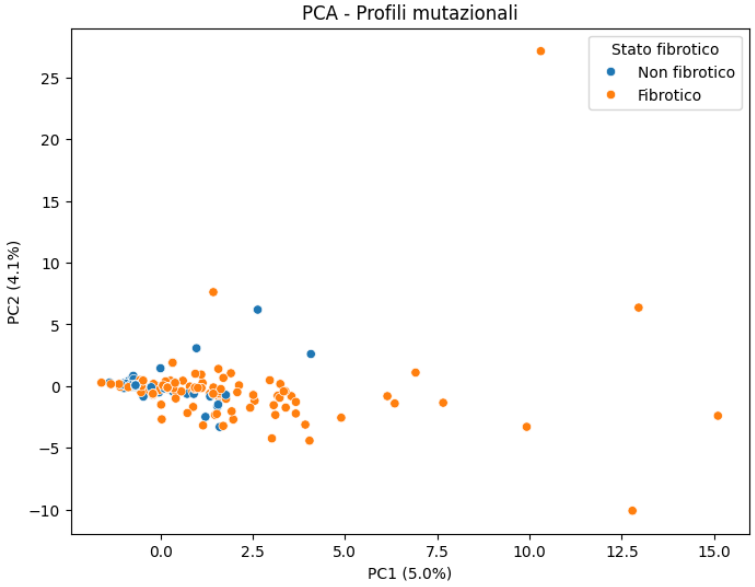

Figura 5.9: Analisi delle componenti principali (PCA) dei profili mutazionali NGS ottenuta escludendo i geni driver principali, colorati per Stato Fibrotico.

Osservando il grafico in Figura 5.9, si nota già una cosa interessante: i pochi pazienti che

si allontanano nettamente dal nucleo centrale della distribuzione sembrano essere prevalentemente fibrotici (arancione). Non c'è una separazione completa tra le due classi, ma i casi più estremi lungo PC1 e PC2 tendono a concentrarsi nella classe fibrotica. Questo potrebbe essere un primo indizio del fatto che alcuni profili mutazionali particolarmente complessi o rari siano associati alla fibrosi, anche se la PCA da sola non mostra una segregazione netta dei due gruppi.

Sono state allora eseguite ulteriori analisi complementari per approfondire la struttura mutazionale con tecniche più adatte a dati genomici sparsi e ad alta dimensionalità. Di seguito sono riportati i grafici.

Nel grafico 5.10 non emerge alcuna separazione evidente tra pazienti JAK2, CALR, MPL o triplo negativi.

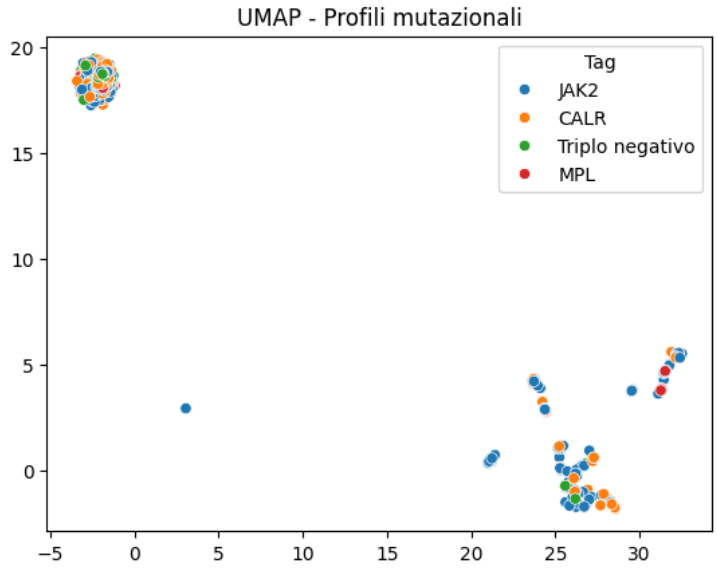

Figura 5.10: UMAP dei profili mutazionali NGS ottenuta escludendo i geni driver principali, colorati per gene driver.

Nel cluster principale (Figura 5.11) situato nella parte superiore sinistra si osserva una predominanza di pazienti non fibrotici, con una distribuzione relativamente mista ma tendenzialmente arricchita di soggetti senza fibrosi. Al contrario, i cluster localizzati nella regione destra del grafico mostrano una maggiore concentrazione di pazienti fibrotici.

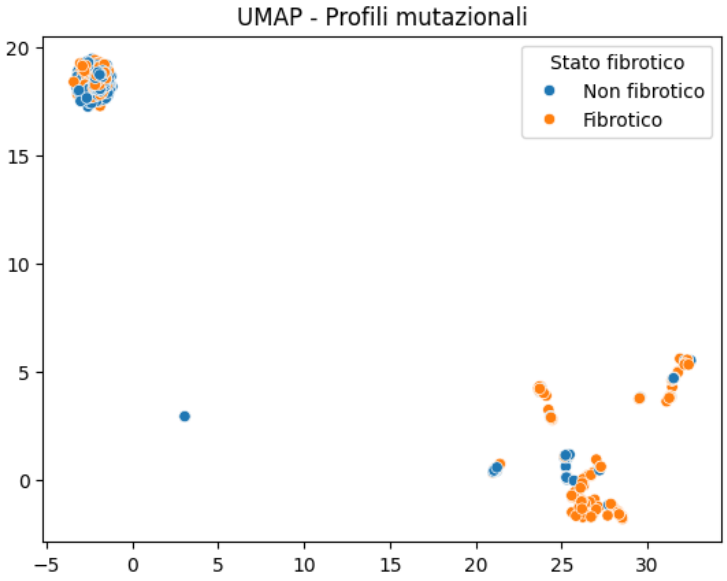

Figura 5.11: UMAP dei profili mutazionali NGS ottenuta escludendo i geni driver principali, colorati per Stato Fibrotico.

Nel Figura 5.12, il t-SNE conferma quanto già osservato sia nella PCA sia nell'UMAP: le mutazioni accessorie non sembrano organizzarsi secondo il gene driver.

La regione superiore del Figura 5.13 (circa y = -5) contiene una prevalenza di pazienti non fibrotici, mentre gran parte delle regioni inferiori è arricchita di pazienti fibrotici. Molti cluster locali risultano costituiti quasi esclusivamente da casi fibrotici. Non si tratta ancora di una separazione perfetta, ma la distribuzione appare decisamente meno casuale rispetto a quella osservata per il gene driver e mostra una segregazione più evidente dei pazienti fibrotici rispetto ai metodi precedenti.

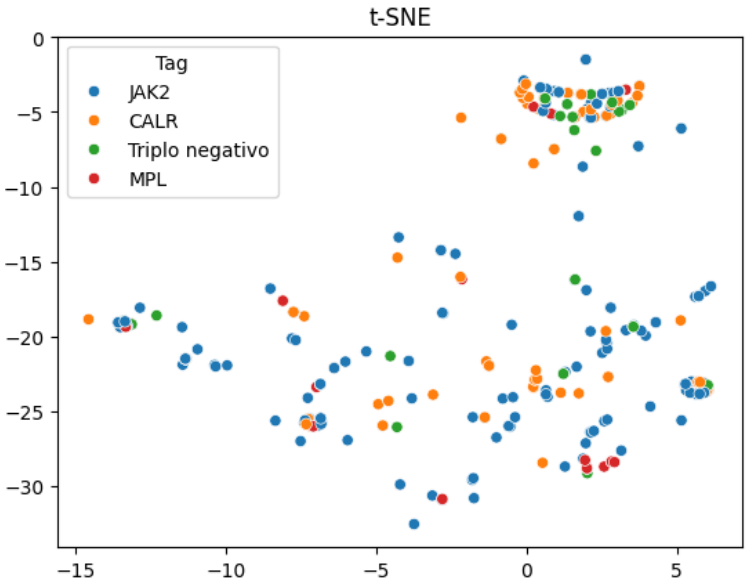

Figura 5.12: t-SNE dei profili mutazionali NGS ottenuta escludendo i geni driver principali, colorati per gene driver.

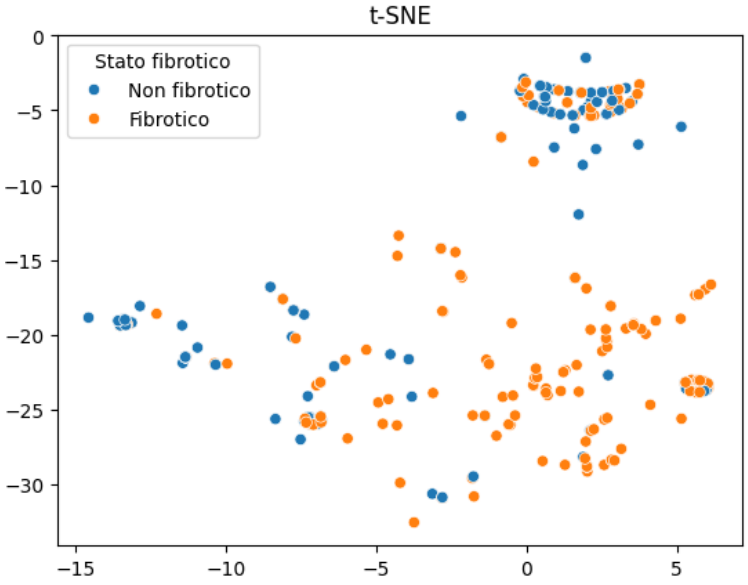

Figura 5.13: t-SNE dei profili mutazionali NGS ottenuta escludendo i geni driver principali, colorati per Stato Fibrotico.

5.3 Risultati dei modelli predittivi

I risultati delle analisi, organizzate in due principali approcci e illustrate nelle Figure 5.14 e 5.21, sono presentati secondo una progressione volta a valutare dapprima la capacità predittiva delle sole informazioni genomiche e, successivamente, il contributo dell'integrazione delle informazioni cliniche e genomiche.

Il primo approccio riguarda l'impiego di modelli di machine learning tradizionali come baseline genomica, mentre il secondo è incentrato sull'utilizzo di TabPFN attraverso una strategia di integrazione progressiva delle diverse tipologie di variabili.

5.3.1 Modelli classici

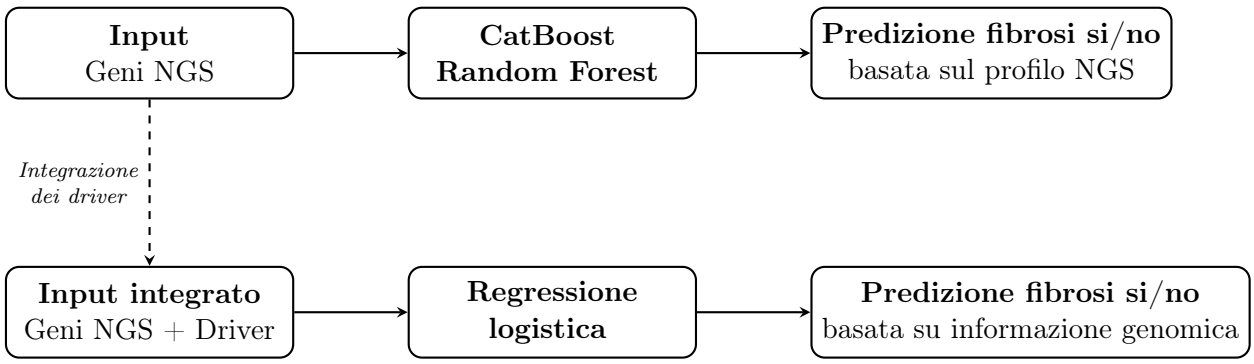

Figura 5.14: Flowchart delle analisi predittive mediante modelli classici utilizzati come baseline genomica.

I risultati dei successivi modelli classici, mostrano le loro prestazioni di classificazione, utilizzando esclusivamente i dati genomici.

Il modello Catboost (Figura 5.15) ha correttamente identificato 45 dei 50 pazienti non fibrotici e 26 dei 57 pazienti fibrotici. La specificità è risultata elevata (90,0%.), indicando una buona capacità di riconoscere i soggetti non fibrotici, mentre la sensibilità per la classe fibrotica è risultata inferiore (45,6%.), evidenziando una tendenza del modello a classificare come non fibrotici una quota rilevante di pazienti affetti da fibrosi (Tabella 5.10).

Tabella 5.10: Prestazioni del modello CatBoost nella classificazione dello stato fibrotico.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.59	0.90	0.71	50
Fibrotico	0.84	0.46	0.59	57

Questi risultati suggeriscono che il modello genera un numero limitato di falsi positivi, ma tende a perdere una parte consistente dei casi fibrotici, come evidenziato dall'elevato numero di falsi negativi (n = 31).

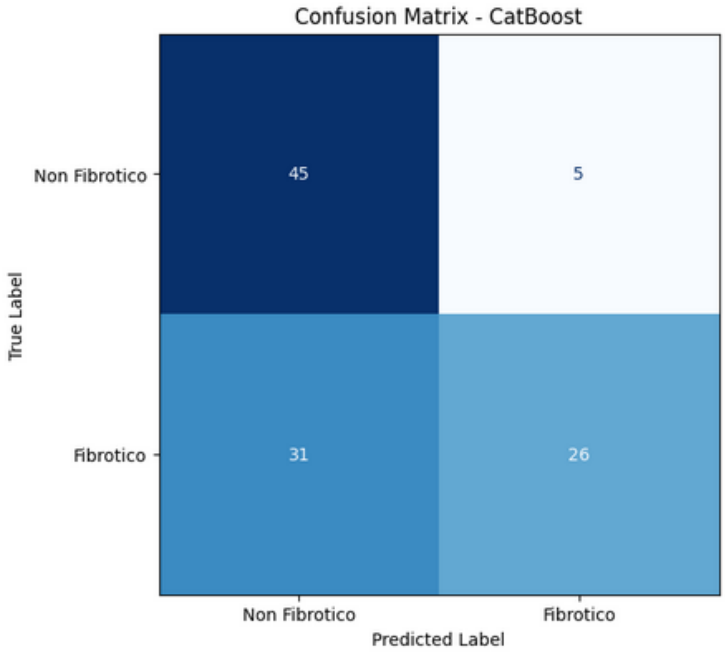

Figura 5.15: Matrice di confusione del modello CatBoost per la classificazione dei pazienti fibrotici e non fibrotici.

Le prestazioni discriminanti del modello sono state ulteriormente valutate mediante l'analisi della curva ROC, riportata in Figura 5.18.

Sebbene le prestazioni siano superiori a quelle attese da una classificazione casuale (AUC = 0.50), il valore osservato evidenzia come le caratteristiche utilizzate dal modello siano soltanto parzialmente in grado di discriminare i due gruppi. Questo risultato è coerente con quanto

osservato nella matrice di confusione, che mostra una buona specificità ma una sensibilità relativamente limitata nei confronti della classe fibrotica.

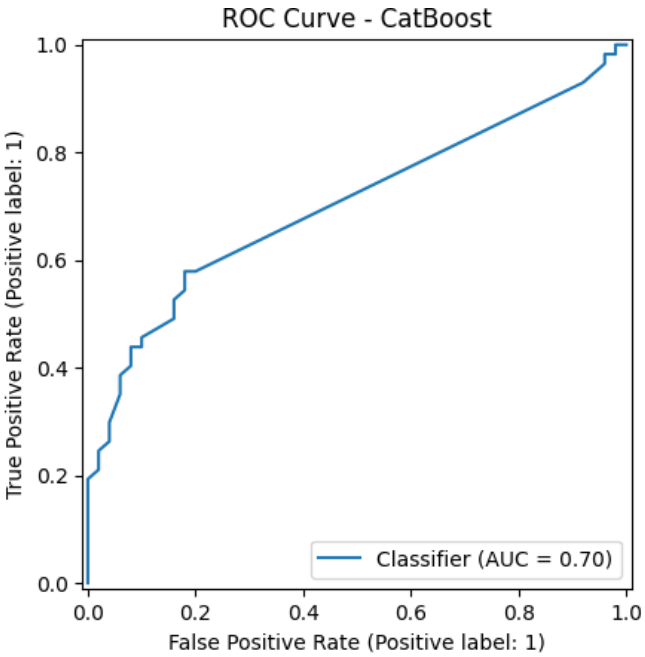

Figura 5.16: Curva ROC del modello CatBoost per la classificazione dei pazienti fibrotici e non fibrotici. La curva mostra il compromesso tra sensibilità (True Positive Rate) e tasso di falsi positivi (False Positive Rate) ai diversi valori di soglia decisionale. L'area sotto la curva (AUC) pari a 0.70 indica una moderata capacità discriminante del modello.

Il modello Random Forest ha mostrato una capacità moderata di discriminare tra pazienti fibrotici e non fibrotici. Come evidenziato dalla matrice di confusione (Figura 5.17), il modello ha classificato correttamente 47 dei 50 pazienti non fibrotici e 23 dei 57 pazienti fibrotici, raggiungendo un'accuratezza complessiva del 65.4

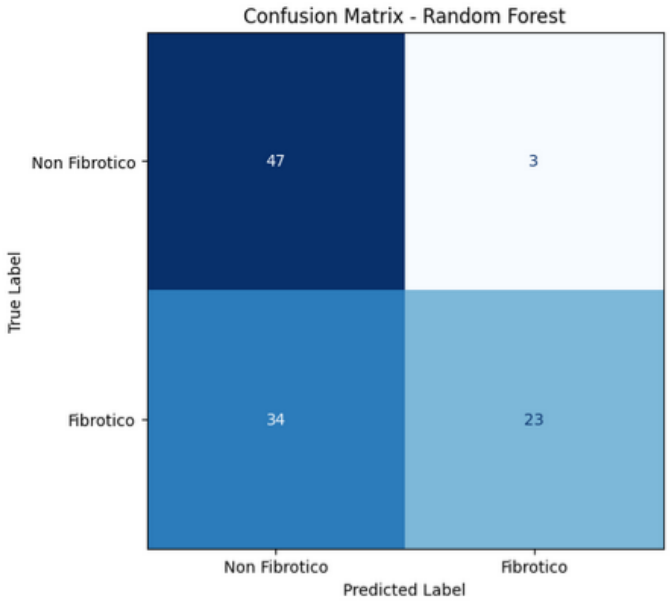

Figura 5.17: Matrice di confusione del modello Random Forest per la classificazione dei pazienti fibrotici e non fibrotici.

Tabella 5.11: Prestazioni del modello Random Forest nella classificazione dello stato fibrotico.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.58	0.94	0.72	50
Fibrotico	0.88	0.40	0.55	57

L'analisi delle prestazioni evidenzia una marcata asimmetria nella capacità di classificazione delle due classi, dimostrando una notevole efficacia nell'identificare correttamente i pazienti non fibrotici e nel limitare il numero di falsi positivi. Al contrario, la sensibilità per la classe fibrotica indica che una quota consistente di pazienti con fibrosi viene erroneamente classificata come non fibrotica.

Questo comportamento suggerisce che il modello tenda ad adottare una strategia classificativa conservativa, privilegiando l'accuratezza nella classificazione dei pazienti non fibrotici a discapito dell'identificazione dei casi fibrotici. Tale tendenza è confermata dall'elevata precisione osservata per la classe fibrotica (0.88), che indica come la maggior parte dei pazienti classificati dal modello come fibrotici presenti effettivamente la condizione, sebbene numerosi casi di fibrosi non vengano riconosciuti.

L'analisi della curva ROC ha restituito un valore di AUC pari a 0.719, indicativo di una capacità discriminante moderata e superiore a quella attesa da una classificazione casuale. Questo risultato suggerisce che il modello sia in grado di cogliere parte del segnale biologico associato allo sviluppo della fibrosi, ma che le variabili utilizzate non consentano una separazione netta tra i due gruppi di pazienti.

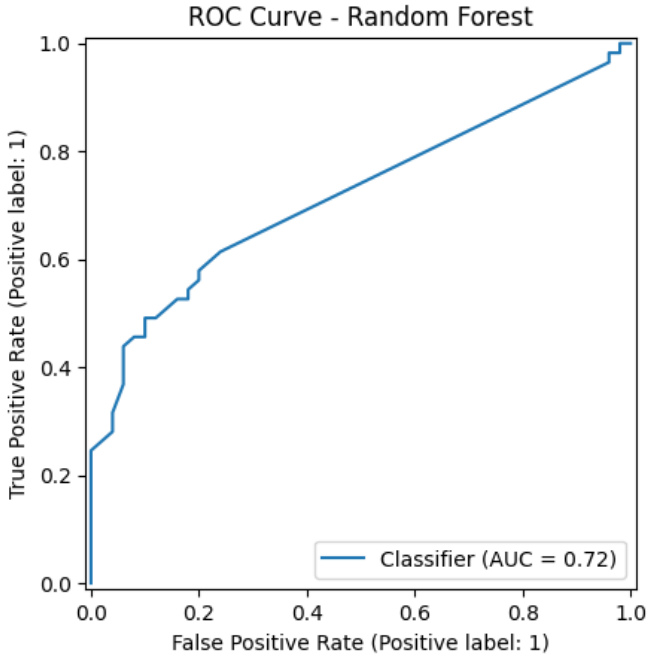

Figura 5.18: Curva ROC del modello Random Forest per la classificazione dei pazienti fibrotici e non fibrotici.

Nel complesso, confrontando le prestazioni del modello Random Forest con quelle del modello CatBoost, emerge un comportamento sostanzialmente simile, caratterizzato da una buona

capacità di identificare i pazienti non fibrotici e da una minore sensibilità nel riconoscimento dei pazienti fibrotici.

La Regressione Logistica è stata addestrata utilizzando sia le informazioni mutazionali ottenute mediante sequenziamento NGS sia il gene driver del paziente.

Come mostrato nella Figura 5.19, il modello ha correttamente classificato 46 dei 50 pazienti non fibrotici e 26 dei 57 pazienti fibrotici, indicando una buona capacità di identificare i pazienti non fibrotici e una capacità moderata di riconoscere i casi di fibrosi.

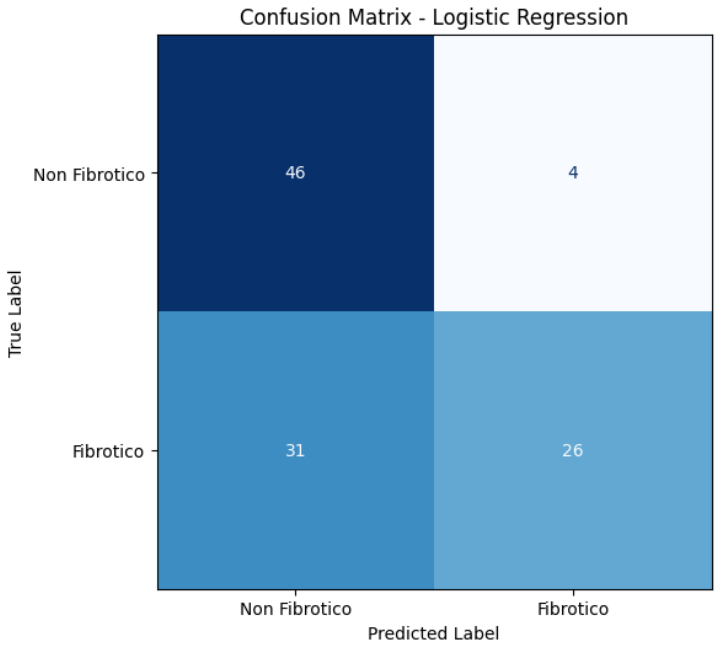

Figura 5.19: Matrice di confusione del modello Regressione Logistica.

Tabella 5.12: Prestazioni del modello Regressione Logistica nella classificazione dello stato fibrotico.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.60	0.92	0.72	50
Fibrotico	0.87	0.46	0.60	57

L'analisi della curva ROC (Figura 5.20) ha evidenziato un valore di AUC pari a 0.75, il più elevato tra i modelli valutati. Tale risultato suggerisce una migliore capacità discriminante

della regressione logistica nel distinguere i pazienti fibrotici da quelli non fibrotici lungo l'intero intervallo delle soglie decisionali.

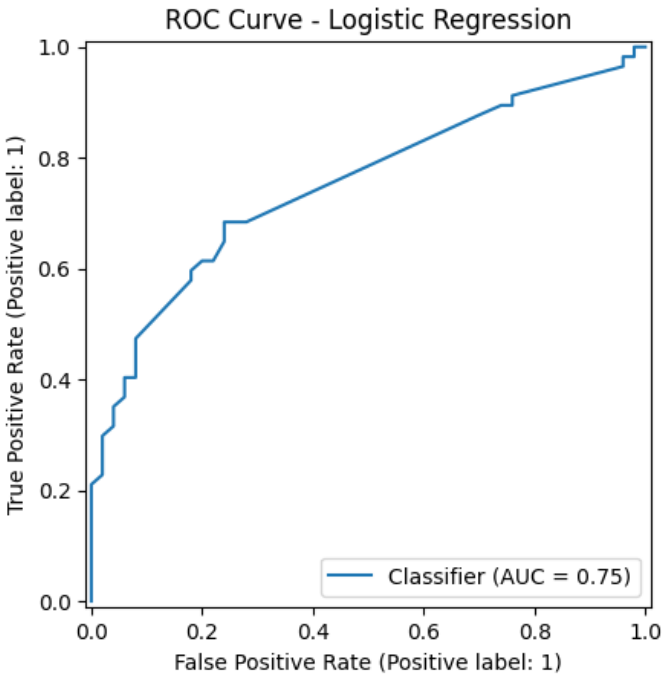

Figura 5.20: Curva ROC del modello Regressione Logistica per la classificazione dei pazienti fibrotici e non fibrotici.

Rispetto a CatBoost e Random Forest, la Regressione Logistica ha mostrato prestazioni complessivamente superiori, ottenendo contemporaneamente la migliore accuratezza e il valore più elevato di AUC. Questo risultato suggerisce che l'informazione associata allo stato fibrotico possa essere descritta efficacemente da relazioni prevalentemente lineari tra le mutazioni accessorie e il gene driver.

Nonostante tali risultati, il numero relativamente elevato di falsi negativi (n = 31) indica che una quota significativa di pazienti fibrotici continua a non essere identificata correttamente dal modello. Ciò suggerisce che la progressione verso la fibrosi sia probabilmente influenzata da ulteriori fattori clinici non completamente rappresentati dalle sole variabili genomiche considerate.

5.3.2 Modello TabPFN

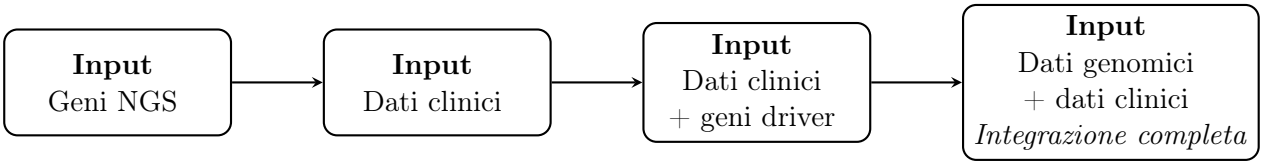

Figura 5.21: Progressiva integrazione delle informazioni genomiche e cliniche come input al modello TabPFN.

Configurazione genomica

Il modello TabPFN addestrato esclusivamente sui dati NGS mostra prestazioni differenti tra le due classi. Come evidenziato dalla matrice di confusione, il modello identifica correttamente 45 campioni non fibrotici su 50, ottenendo un recall pari a 0,90, mentre riconosce correttamente 27 dei 57 campioni fibrotici, con un recall di 0,47.

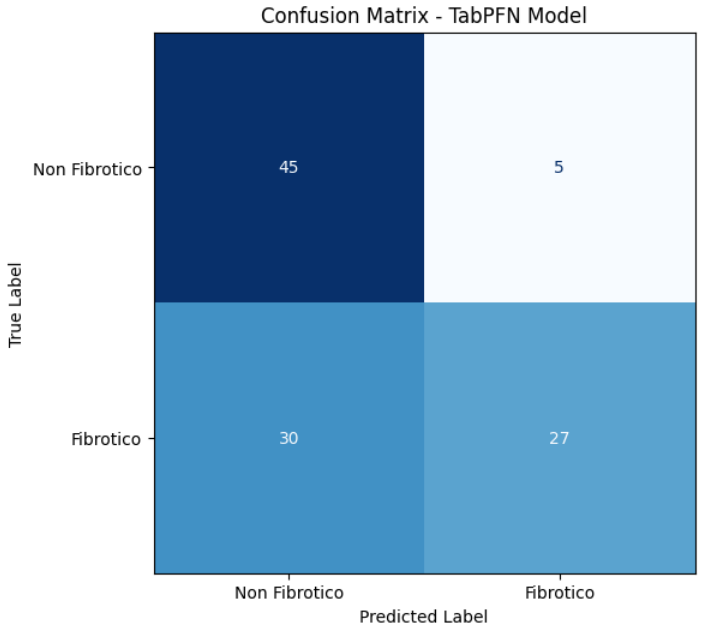

Figura 5.22: Matrice di confusione del modello TabPFN addestrato esclusivamente su geni NGS.

Nel complesso, questi risultati suggeriscono che le sole informazioni derivate dai dati NGS non siano sufficienti a discriminare efficacemente tutti i campioni fibrotici. Il modello sembra

Tabella 5.13: Prestazioni del modello TabPFN basato esclusivamente su dati NGS.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.60	0.90	0.72	50
Fibrotico	0.84	0.47	0.61	57

infatti privilegiare la classificazione come non fibrotico, determinando un numero elevato di falsi negativi.

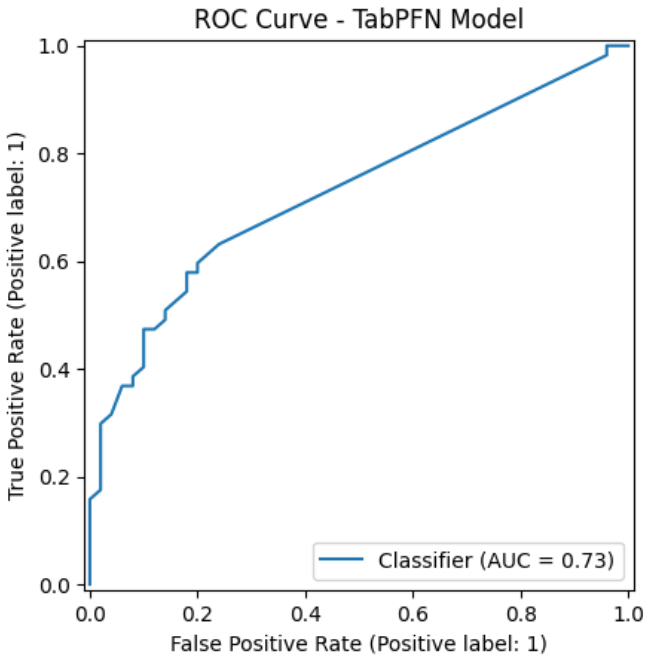

Figura 5.23: Curva ROC del modello TabPFN basato esclusivamente su geni NGS.

Configurazione clinica

Il modello TabPFN basato esclusivamente su dati clinici ha mostrato prestazioni elevate nella classificazione dei pazienti fibrotici e non fibrotici. La sensibilità per la classe fibrotica è risultata pari all'80.7%, mentre la specificità ha raggiunto il 92.0%, indicando una buona capacità del modello sia nel riconoscimento dei pazienti con fibrosi sia nell'esclusione dei pazienti non fibrotici. Il numero limitato di falsi positivi (n = 4) e falsi negativi (n = 11) suggerisce una classificazione complessivamente equilibrata tra le due classi.

Tabella 5.14: Prestazioni del modello TabPFN basato esclusivamente su dati clinici.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.81	0.92	0.86	50
Fibrotico	0.92	0.81	0.86	57

Come illustrato nella Figura 5.24, il classificatore ha correttamente identificato 46 dei 50 pazienti non fibrotici e 46 dei 57 pazienti fibrotici, raggiungendo un'accuratezza complessiva dell'86.0%.

L'analisi della curva ROC (Figura 5.25) ha evidenziato un valore di AUC pari a 0.94, indicativo di un'eccellente capacità discriminante.

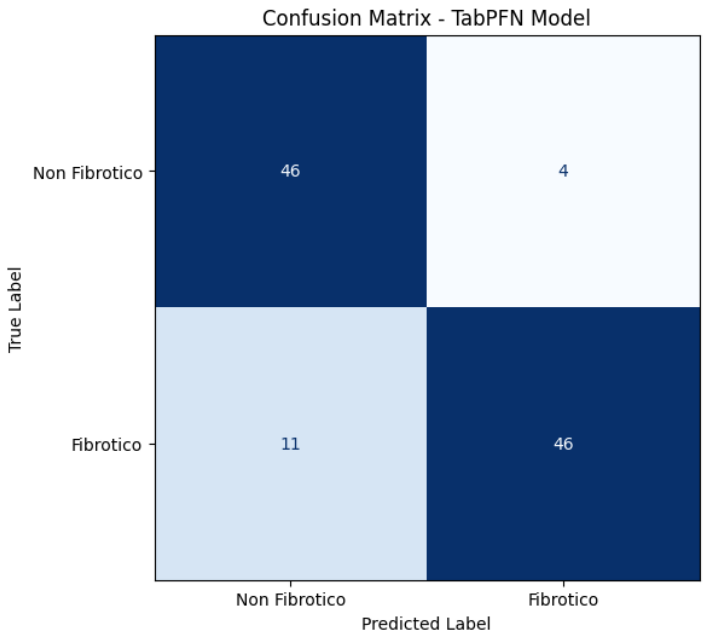

Figura 5.24: Matrice di confusione del modello TabPFN addestrato esclusivamente su variabili cliniche.

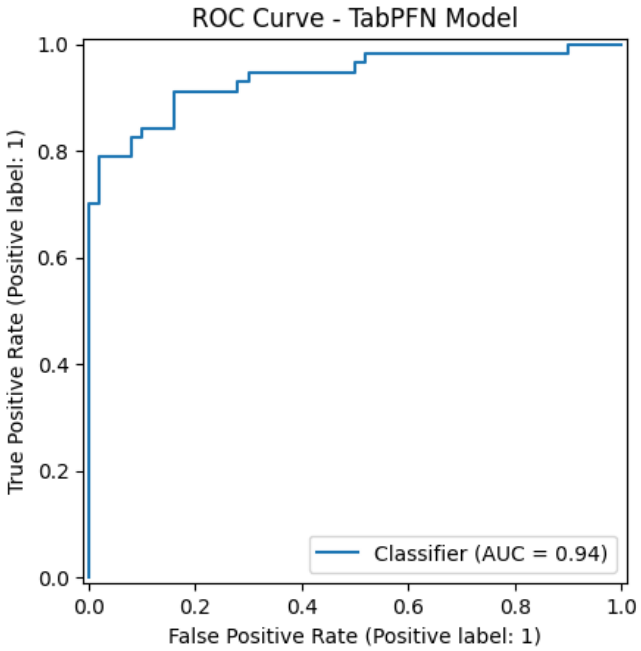

Figura 5.25: Curva ROC del modello TabPFN basato esclusivamente su dati clinici.

Nel complesso, questi risultati suggeriscono che le caratteristiche cliniche e istopatologiche raccolte alla diagnosi contengano una quota rilevante dell'informazione associata allo stato fibrotico. L'integrazione delle informazioni genomiche è finalizzata a valutare se le mutazioni NGS possano apportare un ulteriore contributo predittivo rispetto a un modello già altamente performante basato sulle sole variabili cliniche.

Integrazione del gene driver

Per valutare il contributo specifico del gene driver alla classificazione dello stato fibrotico, è stato sviluppato un modello TabPFN integrando le variabili cliniche con l'informazione relativa allo stato mutazionale driver del paziente (JAK2, CALR, MPL o triplo negativo). L'obiettivo era verificare se tale informazione molecolare, già ampiamente utilizzata nella classificazione biologica delle neoplasie mieloproliferative, fosse in grado di migliorare le prestazioni ottenute utilizzando esclusivamente le variabili cliniche.

Nonostante il modello completo (5.3.2) utilizzi un numero significativamente maggiore di

Tabella 5.15: Prestazioni del modello TabPFN basato su dati clinici e gene driver.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.79	0.96	0.86	50
Fibrotico	0.96	0.77	0.85	57

variabili molecolari, l'aggiunta delle mutazioni accessorie non determina un miglioramento delle prestazioni rispetto all'utilizzo del solo gene driver.

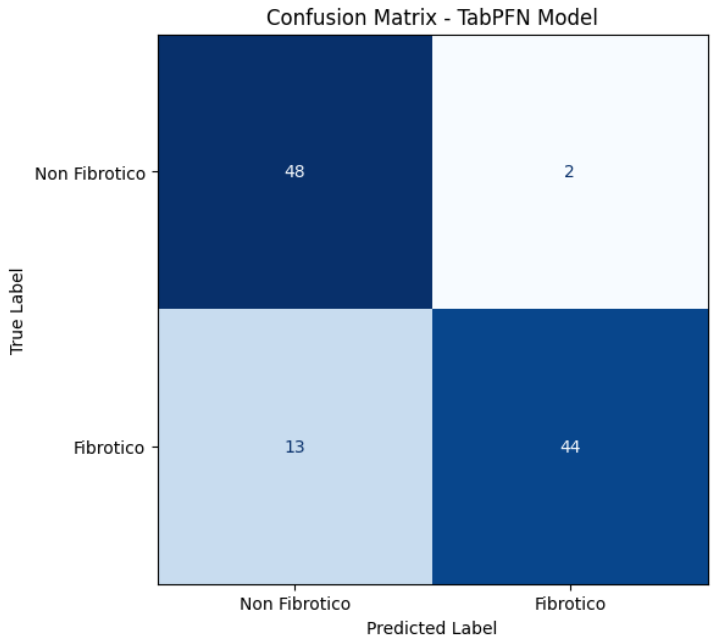

Figura 5.26: Matrice di confusione del modello TabPFN addestrato su variabili cliniche e gene driver

Questi risultati suggeriscono che una quota rilevante dell'informazione genomica associata allo sviluppo della fibrosi sia già catturata dal gene driver principale e dalle caratteristiche cliniche disponibili alla diagnosi. Nella coorte analizzata, le mutazioni NGS accessorie sembrano fornire un contributo incrementale limitato alla classificazione dello stato fibrotico, mentre il gene driver mantiene un ruolo centrale nella caratterizzazione biologica della malattia.

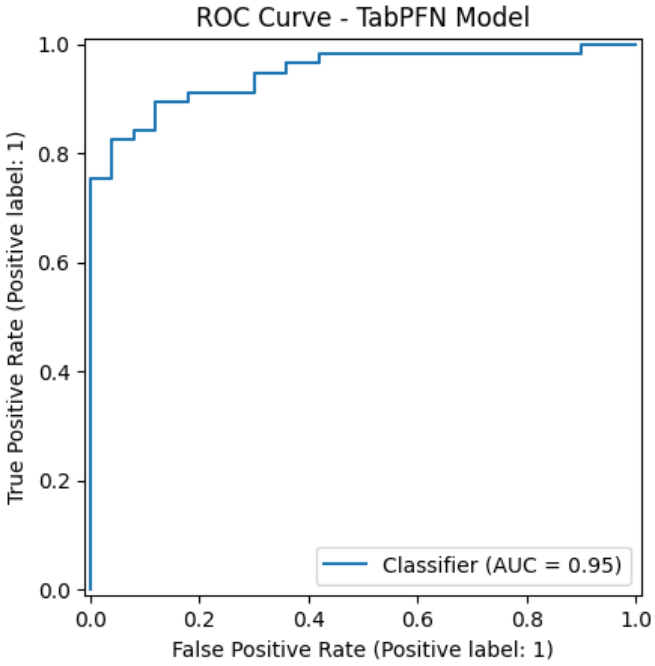

Figura 5.27: Curva ROC del modello TabPFN basato su dati clinici e gene driver.

Integrazione di dati clinici e genomici

In ultima analisi, sono state integrate tutte le informazioni cliniche e genomiche (gene driver e geni NGS) e, come mostrato nella Figura 5.28, il modello ha correttamente classificato 48 dei 50 pazienti non fibrotici e 44 dei 57 pazienti fibrotici, raggiungendo un'accuratezza complessiva dell'86.0%. La specificità è risultata pari al 96.0%, mentre la sensibilità per la classe fibrotica è stata del 77.2%, evidenziando una buona capacità di discriminazione tra le due categorie cliniche.

Tabella 5.16: Prestazioni del modello TabPFN basato su dati clinici e genomici.

Classe	Precisione	Recall	F1-score	Supporto
Non fibrotico	0.79	0.96	0.86	50
Fibrotico	0.96	0.77	0.85	57

L'analisi della curva ROC (Figura 5.29) ha evidenziato un valore di AUC pari a 0.942, indicativo di un'eccellente capacità discriminante. Tuttavia, il confronto con i modelli sviluppati

nelle sezioni precedenti mostra come l'integrazione delle informazioni genomiche non determini un miglioramento sostanziale delle prestazioni rispetto a quanto ottenuto utilizzando esclusivamente le variabili cliniche o la combinazione di dati clinici e gene driver.

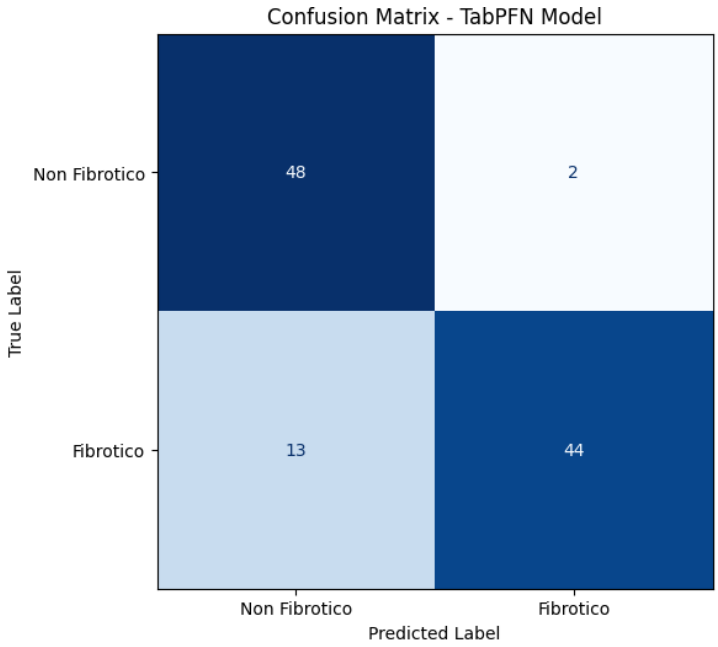

Figura 5.28: Matrice di confusione del modello TabPFN addestrato su variabili cliniche e genomiche

Interessante risulta il contributo della percentuale di mutazione allelica del gene driver JAK2. Nonostante essa sia frequentemente associata alla progressione della malattia e all'evoluzione fibrotica, la sua inclusione nel modello non ha prodotto un incremento della capacità discriminante.

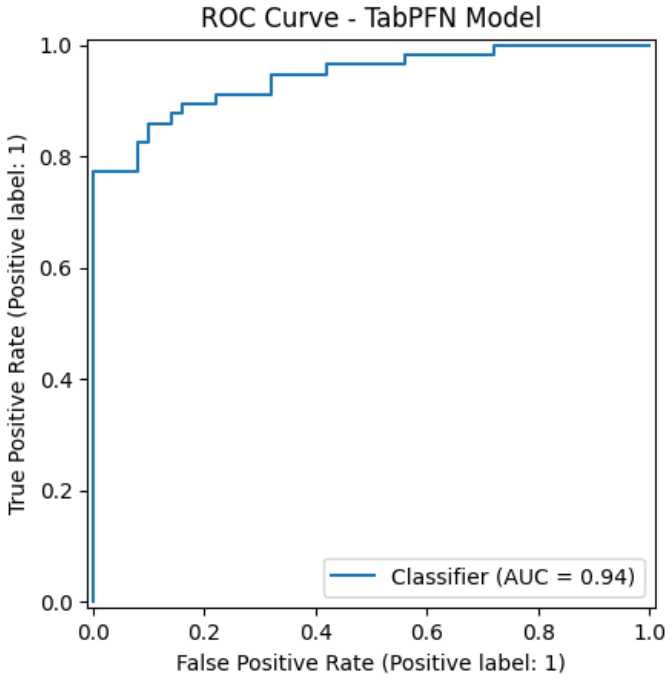

Figura 5.29: Curva ROC del modello TabPFN basato su dati clinici e genomici.

Nel complesso, questi risultati suggeriscono che gran parte dell'informazione biologica associata allo stato fibrotico sia già catturata dalle sole caratteristiche cliniche.

5.4 Confronto delle performance

Per ciascuna strategia di adattamento dei pesi nei Tabular Foundation Models nelle configurazioni specificate, sono state valutate le prestazioni sul task di classificazione del fenotipo fibrotico mediante le metriche definite in precedenza (sezione 4.8). I risultati medi ottenuti sui diversi fold per ogni modello sono riportati nella Tabella 5.17.

Tabella 5.17: Risultati del Benchmark con modelli TabTune e strategie di tuning diverse.

Model	Strategy	Accuracy	F1 Score	ROC AUC
TabPFN	inference	0.887180	0.887244	0.959274
TabPFN	base-ft	0.887180	0.887244	0.959274
TabICL	peft	0.883354	0.882875	0.960786
TabICL	base-ft	0.881485	0.881352	0.962615
TabICL	inference	0.881485	0.881367	0.961104
Mitra	peft	0.557997	0.420768	0.858703
Mitra	inference	0.512890	0.350331	0.582277

Sulla base delle configurazioni usate nel benchmarking, per ogni modello sono stati effettuati esperimenti sia utilizzando esclusivamente le caratteristiche originali del dataset sia sfruttando gli embedding pre-addestrati `Mut2Vec` sulle prestazioni finali. Nella Tabella 5.18 e 5.19 sono riportati i valori medi ± deviazione standard.

Tabella 5.18: Risultati medi della 10-fold Cross-Validation senza embedding pre-addestrati di `Mut2Vec`.

Model	Strategy	Accuracy	F1 Score	ROC AUC
TabPFN	base-ft	0.887 ± 0.064	0.889 ± 0.062	0.964 ± 0.033
TabPFN	inference	0.887 ± 0.064	0.889 ± 0.062	0.964 ± 0.033
TabICL	inference	0.895 ± 0.046	0.896 ± 0.045	0.961 ± 0.039
TabICL	base-ft	0.891 ± 0.051	0.892 ± 0.050	0.962 ± 0.039
TabICL	peft	0.887 ± 0.064	0.889 ± 0.062	0.961 ± 0.038
Mitra	peft	0.564 ± 0.201	0.446 ± 0.272	0.863 ± 0.094
Mitra	inference	0.551 ± 0.161	0.402 ± 0.185	0.504 ± 0.386

Tabella 5.19: Risultati medi della 10-Fold Cross-Validation con embedding pre-addestrati di `Mut2Vec`. Tra parentesi è indicato il p-value corretto con il metodo di Holm ottenuto mediante il test di Wilcoxon signed-rank; i p-value significativi ($\alpha = 0.05$) sono evidenziati in grassetto.

Model	Strategy	Accuracy	F1 Score	ROC AUC
TabPFN	base-ft	0.895 ± 0.055	0.896 ± 0.052 (0.656)	0.966 ± 0.033 (1.000)
TabPFN	inference	0.895 ± 0.055	0.896 ± 0.052 (0.656)	0.966 ± 0.033 (1.000)
TabICL	inference	0.848 ± 0.050	0.850 ± 0.047 (**0.014**)	0.926 ± 0.043 (**0.014**)
TabICL	base-ft	0.844 ± 0.051	0.846 ± 0.050 (**0.023**)	0.927 ± 0.044 (**0.020**)
TabICL	peft	0.855 ± 0.049	0.857 ± 0.046 (0.336)	0.941 ± 0.040 (**0.020**)
Mitra	peft	0.660 ± 0.128	0.649 ± 0.183 (0.098)	0.743 ± 0.106 (**0.039**)
Mitra	inference	0.590 ± 0.163	0.473 ± 0.218 (0.656)	0.705 ± 0.174 (0.393)

Il confronto tra i risultati ottenuti evidenzia un comportamento differente tra i modelli considerati. Dai risultati riportati nella Tabella 5.18 e Tabella 5.19 emerge che `TabICL` ha le migliori prestazioni in assenza degli embedding; l'introduzione degli embedding comporta invece una diminuzione delle sue prestazioni in tutte le strategie considerate.

Al contrario, `TabPFN` mostra un lieve miglioramento con l'utilizzo degli embedding, indicando il fatto che `TabPFN` riesce a sfruttare in misura maggiore l'informazione semantica catturata dagli embedding di `Mut2Vec`. Infine, `Mitra` rimane il modello meno competitivo in entrambe le configurazioni.

Per verificare se tali differenze fossero statisticamente significative, è stato applicato il test non parametrico di Wilcoxon signed-rank sui risultati della 10-Fold-Cross Validation, confrontando, per ciascun modello e strategia di addestramento (Tabella 5.18 e Tabella 5.19), le prestazioni ottenute con e senza embedding. I p-value sono stati successivamente corretti mediante il metodo di Holm per controllare l'errore dovuto ai confronti multipli (Appendice A).

Dunque, gli embedding consentono un miglioramento rispetto alla versione senza embedding, ma le prestazioni di `Mitra` rimangono significativamente inferiori rispetto a quelle ottenute da `TabICL` e `TabPFN`, indicando una minore capacità del modello di generalizzare sul dataset considerato (Tabella 5.20 e Tabella 5.21).

L'ipotesi metodologica è che il divario prestazionale osservato può essere interpretato alla luce della diversa predisposizione dei modelli al dominio tabulare. Nel dataset considerato, la sola introduzione di una rappresentazione più ricca delle variabili mediante embedding migliora le prestazioni di Mitra, suggerendo che la rappresentazione degli input costituisca almeno in parte un fattore rilevante, ma non è l'unico. La differenza è riconducibile principalmente alla diversa capacità dei modelli di sfruttare la struttura del dato tabulare e di apprendere le relazioni rilevanti a partire dai dati disponibili.

Tabella 5.20: Confronto fra modelli senza embedding. Risultati Wilcoxon signed-rank test con correzione di Holm per F1 Score e ROC AUC.

Model 1	**Strategy 1**	**Model 2**	**Strategy 2**	**F1 Score**		**ROC AUC**	
				p	**Adj.** p	p	**Adj.** p
Mitra	inference	Mitra	peft	0.6875	1.0000	0.0059	0.0645
Mitra	inference	TabICL	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabICL	inference	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabICL	peft	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabPFN	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabPFN	inference	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabICL	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabICL	inference	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabICL	peft	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabPFN	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabPFN	inference	0.0020	0.0410	0.0020	0.0410
TabICL	base-ft	TabICL	inference	1.0000	1.0000	0.5625	1.0000
TabICL	base-ft	TabICL	peft	0.8125	1.0000	0.6875	1.0000
TabICL	base-ft	TabPFN	base-ft	0.8125	1.0000	0.6250	1.0000
TabICL	base-ft	TabPFN	inference	0.8125	1.0000	0.6250	1.0000
TabICL	inference	TabICL	peft	0.4688	1.0000	0.4375	1.0000
TabICL	inference	TabPFN	base-ft	0.4688	1.0000	0.4316	1.0000
TabICL	inference	TabPFN	inference	0.4688	1.0000	0.4316	1.0000
TabICL	peft	TabPFN	base-ft	0.8457	1.0000	0.5566	1.0000
TabICL	peft	TabPFN	inference	0.8457	1.0000	0.5566	1.0000
TabPFN	base-ft	TabPFN	inference	1.0000	1.0000	1.0000	1.0000

Tabella 5.21: Confronto fra modelli con embedding. Risultati Wilcoxon signed-rank test con correzione di Holm per F1 Score e ROC AUC.

Model 1	**Strategy 1**	**Model 2**	**Strategy 2**	**F1 Score**		**ROC AUC**	
				p	**Adj.** p	p	**Adj.** p
Mitra	inference	Mitra	peft	0.0645	0.3223	0.1055	0.3164
Mitra	inference	TabICL	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabICL	inference	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabICL	peft	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabPFN	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	inference	TabPFN	inference	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabICL	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabICL	inference	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabICL	peft	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabPFN	base-ft	0.0020	0.0410	0.0020	0.0410
Mitra	peft	TabPFN	inference	0.0020	0.0410	0.0020	0.0410
TabICL	base-ft	TabICL	inference	0.5781	1.0000	0.5703	1.0000
TabICL	base-ft	TabICL	peft	0.4609	1.0000	0.0020	0.0410
TabICL	base-ft	TabPFN	base-ft	0.0020	0.0410	0.0020	0.0410
TabICL	base-ft	TabPFN	inference	0.0020	0.0410	0.0020	0.0410
TabICL	inference	TabICL	peft	0.4609	1.0000	0.0039	0.0410
TabICL	inference	TabPFN	base-ft	0.0020	0.0410	0.0020	0.0410
TabICL	inference	TabPFN	inference	0.0020	0.0410	0.0020	0.0410
TabICL	peft	TabPFN	base-ft	0.0039	0.0410	0.0020	0.0410
TabICL	peft	TabPFN	inference	0.0039	0.0410	0.0020	0.0410
TabPFN	base-ft	TabPFN	inference	1.0000	1.0000	1.0000	1.0000

Di seguito, sono riportate le Tabelle che mostrano le performance dei modelli considerati precedentemente, con stessa configurazione, ma addestrati escludendo le informazioni derivanti dalla biospia osteomidollare.

Tabella 5.22: Risultati medi della 10-fold Cross-Validation senza embedding pre-addestrati di `Mut2Vec` e senza l'inclusione delle variabili relative alla biopsia ossea.

Model	**Strategy**	**Accuracy**	**F1 Score**	**ROC AUC**
TabPFN	base-ft	0.833 ± 0.057	0.836 ± 0.052	0.941 ± 0.040
TabPFN	inference	0.833 ± 0.057	0.836 ± 0.052	0.941 ± 0.040
TabICL	inference	0.831 ± 0.071	0.833 ± 0.067	0.934 ± 0.045
TabICL	base-ft	0.827 ± 0.070	0.830 ± 0.067	0.934 ± 0.048
TabICL	peft	0.835 ± 0.065	0.837 ± 0.062	0.933 ± 0.047
Mitra	peft	0.541 ± 0.150	0.396 ± 0.179	0.797 ± 0.072
Mitra	inference	0.515 ± 0.152	0.371 ± 0.181	0.589 ± 0.228

Il confronto tra le configurazioni con e senza le variabili relative alla biopsia ossea evidenzia una riduzione delle prestazioni complessive dei modelli in seguito alla loro esclusione. In assenza di embedding `Mut2Vec`, TabPFN mostra una riduzione dell'Accuracy da 0.887 a

0.833 e del F1 Score da 0.889 a 0.836, mentre la ROC AUC passa da 0.964 a 0.941. Un andamento analogo si osserva per TabICL.

L'introduzione degli embedding Mut2Vec non modifica sostanzialmente questo andamento. Per TabPFN, l'esclusione della biopsia ossea porta a una lieve riduzione dell'Accuracy (da 0.895 a 0.842) e del F1 Score (da 0.896 a 0.845), mentre la ROC AUC passa da 0.966 a 0.936. Per TabICL si osserva una riduzione più marcata delle prestazioni.

Tabella 5.23: Risultati medi della 10-fold Cross-Validation con embedding pre-addestrati di `Mut2Vec` e senza l'inclusione delle variabili relative alla biopsia ossea. Tra parentesi è indicato il p-value corretto con il metodo di Holm ottenuto mediante il test di Wilcoxon signed-rank; i p-value significativi ($\alpha = 0.05$) sono evidenziati in grassetto.

Model	Strategy	Accuracy	F1 Score	ROC AUC
TabPFN	base-ft	0.842 ± 0.045	0.845 ± 0.042 (1.000)	0.936 ± 0.038 (1.000)
TabPFN	inference	0.842 ± 0.045	0.845 ± 0.042 (1.000)	0.936 ± 0.038 (1.000)
TabICL	inference	0.780 ± 0.068	0.785 ± 0.063 (0.058)	0.881 ± 0.049 (**0.013**)
TabICL	base-ft	0.780 ± 0.069	0.785 ± 0.063 (0.136)	0.883 ± 0.047 (**0.014**)
TabICL	peft	0.793 ± 0.066	0.797 ± 0.061 (0.136)	0.894 ± 0.050 (**0.014**)
Mitra	peft	0.647 ± 0.134	0.635 ± 0.181 (**0.027**)	0.728 ± 0.072 (0.109)
Mitra	inference	0.455 ± 0.142	0.296 ± 0.150 (1.000)	0.490 ± 0.208 (1.000)

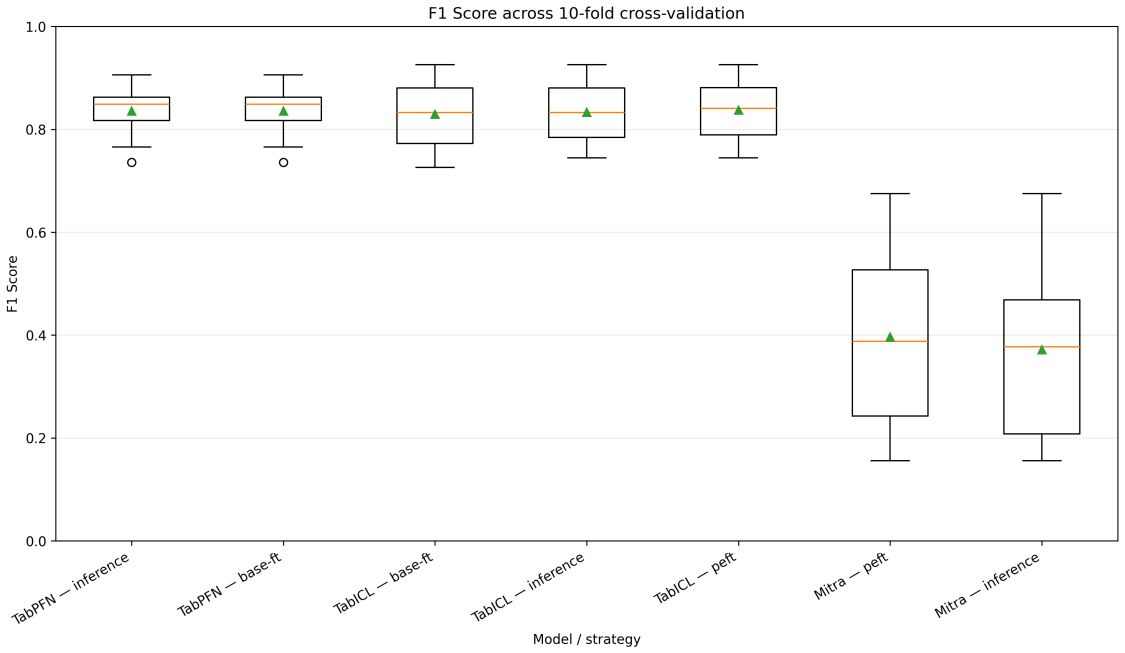

Figura 5.30: Boxplot dei valori di F1 Score ottenuti nei 10 fold della cross-validation per i sette modelli e strategie considerate.

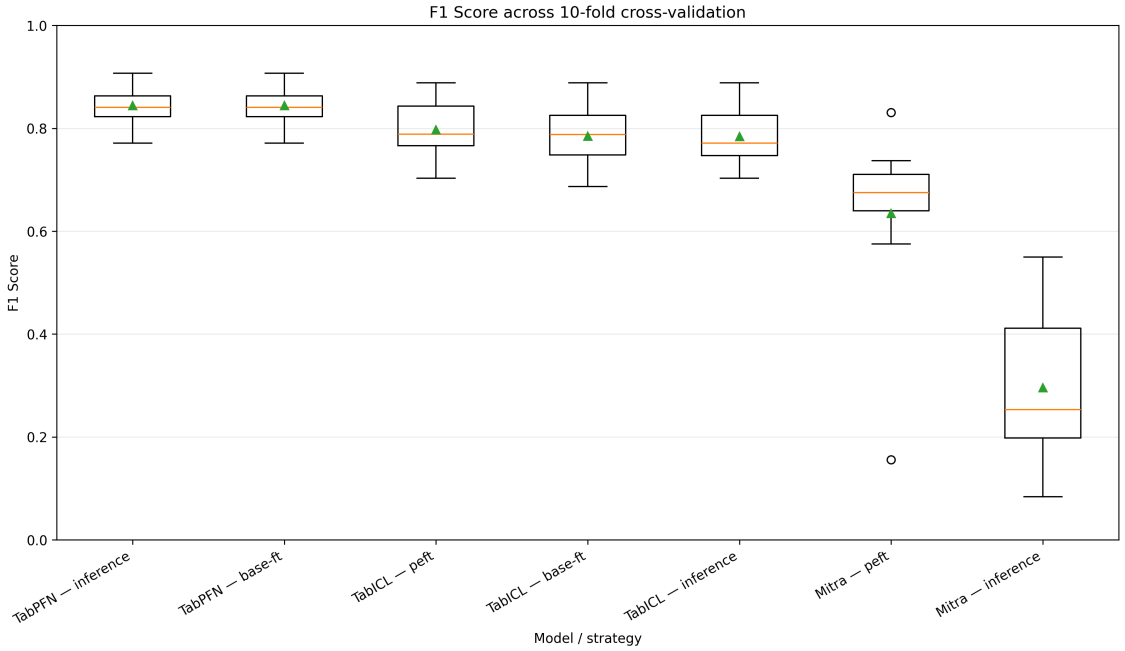

Figura 5.31: Boxplot dei valori di F1 Score ottenuti nei 10 fold della cross-validation per i sette modelli e strategie considerate, con embedding

5.5 Fine-tuning

Come descritto nella sezione 4.7, l'ottimizzazione è stata effettuata mediante grid search.

Tabella 5.24: Risultati dell'ottimizzazione degli iperparametri mediante grid search

Trial	**n_estimators**	**Temperature**	**Epochs**	**Learning Rate**	**Accuracy**	**ROC AUC**
1	8	0.5	1	1×10^{-5}	0.895	0.966
2	8	0.5	1	2×10^{-5}	0.895	0.966
⋮	⋮	⋮	⋮	⋮	⋮	⋮
80	32	1.5	5	2×10^{-5}	0.895	0.966
81	32	1.5	5	5×10^{-5}	0.895	0.966

L'analisi dei risultati (Tabella 5.24) l'esito della procedura di ottimizzazione degli iperparametri, condotta valutando complessivamente 81 configurazioni differenti. La ricerca ha considerato la combinazione di diversi valori di `n_estimators`, `temperature`, numero di `epochs` e `learning rate`, al fine di individuare la configurazione in grado di massimizzare le prestazioni predittive del modello.

Si evidenzia che tutte le combinazioni di iperparametri producono gli stessi valori di Accuracy e ROC AUC, indicando che le prestazioni del modello TabPFN risultano sostanzialmente insensibili alle variazioni degli iperparametri considerati. Questo comportamento suggerisce che, nel contesto del dataset utilizzato, il modello raggiunge prestazioni stabili indipendentemente dalla configurazione adottata.

5.6 Risultati della Permutation Feature Importance

La prima analisi di Permutation Feature Importance, condotta sull'intero set di feature, ha evidenziato una netta prevalenza della variabile relativa al grado di fibrosi, che presenta un valore di importance pari a 0.0584. Tra le variabili successive per importanza compaiono sia caratteristiche cliniche, come l'ematocrito (0.0034), sia feature relative alla gestione dei dati mancanti. Anche la variabile genomica `EZH2` presenta un valore di importance pari a 0.0015. La presenza, tra le feature maggiormente rilevanti, di variabili legate alla missingness ha reso meno immediata l'interpretazione del contributo delle singole caratteristiche cliniche e genomiche. Alla luce di questi risultati, è stata quindi effettuata la successiva fase di feature selection, finalizzata a ottenere un insieme di variabili più selettivo e maggiormente interpretabile.

Tabella 5.25: Top 5 feature per Permutation feature importance.

Feature	**Importance**	**Std**
BOMFibrosi	0.0584	0.0348
Eritropoietina_missing	0.0046	0.0038
Ematocrito	0.0034	0.0089
LDH_lattatoDeidrogenasi__missing	0.0015	0.0031
EZH2	0.0015	0.0031

Anche i risultati della Forward Stepwise Feature Selection, sull'intero insieme di variabili del dataset, mostrano come il punteggio F1 medio, utilizzando la sola variabile relativa al grado di fibrosi, raggiunge già lo 0.9012 e, l'inclusione di ulteriori variabili non ha prodotto miglioramenti apprezzabili delle prestazioni.

Tabella 5.26: Risultati della forward stepwise selection.

Step	**Selected Features**	**Mean F1**	**Δ F1**
1	BOMFibrosi	0.9012	inf
2	BOMFibrosi, CD34Circolanti	0.9129	0.0117
3	BOMFibrosi, CD34Circolanti, SRSF2	0.9152	0.0023
4	BOMFibrosi, CD34Circolanti, SRSF2, BiopsiaOssea	0.9152	0.0000
5	BOMFibrosi, CD34Circolanti, SRSF2, BiopsiaOssea, JAK2V617F_	0.9152	0.0000

Tabella 5.27: Top 5 feature per Permutation Feature importance sul modello addestrato sulle features selezionate.

Feature	**Importance**	**Std**
BOMFibrosi	0.2041	0.0475
CD34Circolanti	0.0166	0.0164
JAK2V617F_	0.0079	0.0105
BiopsiaOssea	0.0000	0.0000
SRSF2	-0.0119	0.0155

In assenza delle informazioni relative alla biopsia ossea, la Permutation Feature Importance evidenzia come variabile maggiormente rilevante la dimensione della milza, con un valore di importance pari a 0.0435, seguita dalla conta dei CD34 circolanti (0.0184) e dall'emoglobina (0.0152).

Tabella 5.28: Top 5 feature per Permutation feature importance senza considerare informazioni relative alla biopsia ossea.

Feature	**Importance**	**Std**
Milza	0.043485	0.028223
CD34Circolanti	0.018378	0.023391
Emoglobina	0.015177	0.017756
LDH_lattatoDeidrogenasi_	0.014217	0.023617
Piastrine	0.011714	0.024566

Tabella 5.29: Risultati della Forward Stepwise Feature Selection dopo l'esclusione delle variabili relative alla biopsia osteomidollare.

Step	**Feature selezionate**	**Mean F1**	**ΔF1**
1	Milza	0.7577	inf
2	Milza, CD34Circolanti	0.8197	0.0620
3	Milza, CD34Circolanti, LDH	0.8369	0.0172
4	Milza, CD34Circolanti, LDH, Eritropoietina	0.8487	0.0118
5	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina	0.8583	0.0096
6	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F	0.8679	0.0096
7	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F, EZH2	0.8701	0.0022
8	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F, EZH2, NF1	0.8749	0.0048
9	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F, EZH2, NF1, Tag	0.8773	0.0024
10	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F, EZH2, NF1, Tag, SRSF2	0.8822	0.0049
11	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F, EZH2, NF1, Tag, SRSF2, ASXL1	0.8822	-0.0001
12	Milza, CD34Circolanti, LDH, Eritropoietina, Emoglobina, JAK2V617F, EZH2, NF1, Tag, SRSF2, ASXL1, CALR	0.8822	-0.0000

Si osserva, inoltre, un comportamento tipico della forward stepwise:

- Nelle prime iterazioni della procedura di Forward Stepwise Feature Selection si osserva un incremento consistente del punteggio F1, indicando che le prime feature selezionate apportano un contributo informativo rilevante.
- Con l'aggiunta di ulteriori variabili, i miglioramenti delle prestazioni diventano progressivamente più contenuti, evidenziando una progressiva saturazione dell'informazione disponibile.
- Dopo circa dieci feature, la curva delle prestazioni entra in una fase di *plateau*: ulteriori feature non determinano incrementi significativi del punteggio F1 e, in alcuni casi, producono leggere oscillazioni attribuibili alla variabilità della validazione incrociata.

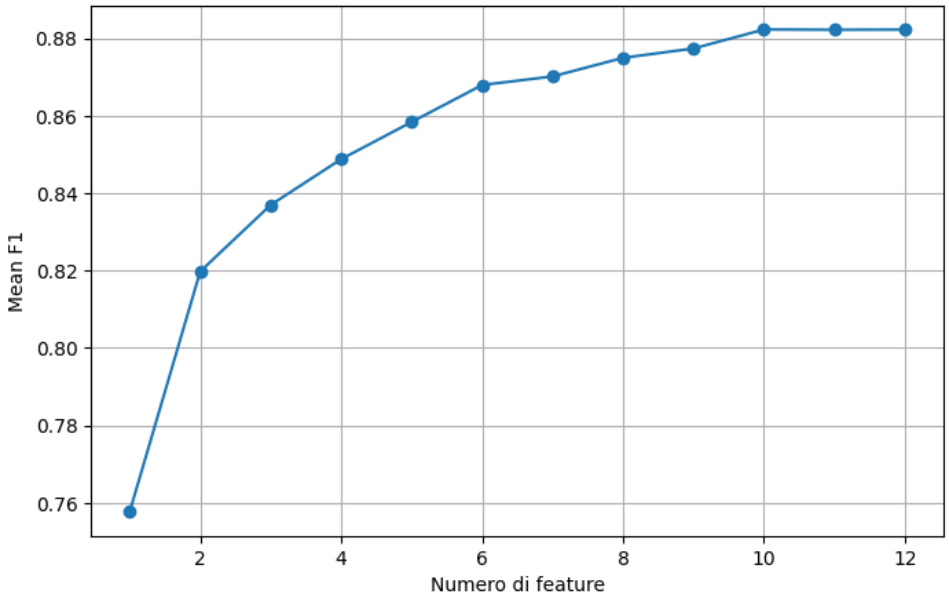

Figura 5.32: Andamento del punteggio F1 medio ottenuto mediante validazione incrociata durante la procedura di Forward Stepwise Feature Selection, in funzione del numero di feature selezionate.

Tabella 5.30: Permutation Feature Importance sul modello addestrato sulle feature selezionate (5.29).

Feature	**Importance**	**Std**
Milza	0.0634	0.0267
Emoglobina	0.0518	0.0344
CD34Circolanti	0.0344	0.0211
JAK2V617F_	0.0205	0.0097
LDH_lattatoDeidrogenasi_	0.0168	0.0256
Tag	0.0095	0.0165
EZH2	0.0018	0.0080
Eritropoietina	-0.0041	0.0151
NF1	-0.0069	0.0092
ASXL1	-0.0071	0.0082
SRSF2	-0.0172	0.0169

La PFI (Tabella 5.30) ha evidenziato che le variabili maggiormente influenti nel modello finale sono risultate Milza, Emoglobina e CD34Circolanti; al contrario, alcune feature se-

lezionate dalla procedura di Forward Stepwise, quali `ASXL1`, `NF1` e `SRSF2`, hanno mostrato un'importanza prossima allo zero o lievemente negativa.

Tale risultato non è in contraddizione con la procedura di selezione delle feature. La Forward Stepwise Selection valuta infatti il contributo incrementale di ciascuna variabile durante la costruzione del modello, mentre la Permutation Feature Importance misura l'effetto della rimozione dell'informazione fornita da una feature nel modello già addestrato. Di conseguenza, una variabile può risultare utile durante la fase di selezione pur fornendo un contributo marginale nel modello finale, soprattutto in presenza di correlazioni o ridondanze tra le feature.

Tabella 5.31: Risultati della Forward Stepwise Feature Selection usando solo features mutazionali.

Step	**Selected features**	**Mean F1**	**ΔF1**
1	ASXL1	0.5913	–
2	ASXL1, SRSF2	0.6318	0.0405
3	ASXL1, SRSF2, SF3B1	0.6499	0.0181
4	ASXL1, SRSF2, SF3B1, KMT2A	0.6660	0.0162
5	ASXL1, SRSF2, SF3B1, KMT2A, EZH2	0.6806	0.0145
6	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1	0.6914	0.0108
7	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3	0.7003	0.0089
8	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8	0.7055	0.0052
9	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8, ZRSR2	0.7110	0.0055
10	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8, ZRSR2, CUX1	0.7161	0.0051
11	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8, ZRSR2, CUX1, NRAS	0.7186	0.0025
12	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8, ZRSR2, CUX1, NRAS, CBL	0.7217	0.0031
13	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8, ZRSR2, CUX1, NRAS, CBL, SUZ12	0.7242	0.0025
14	ASXL1, SRSF2, SF3B1, KMT2A, EZH2, U2AF1, SH2B3, PRPF8, ZRSR2, CUX1, NRAS, CBL, SUZ12, RUNX2	0.7242	0.0000

La feature selection eseguita sul modello addestrato esclusivamente su dati genomici, mostra

quali mutazioni siano maggiormente associate alla classificazione, quanta informazione sia contenuta nel solo profilo mutazionale e se sia possibile discriminare le classi utilizzando esclusivamente dati genomici.

La variabile che identifica il gene driver del paziente, non è stata selezionata dalla procedura di feature selection, non apportando informazione aggiuntiva rispetto a quella già fornita dalle variabili genetiche e, di conseguenza, la sua inclusione non determina un miglioramento delle prestazioni predittive misurate tramite l'F1-score. Questo risultato è verosimilmente dovuto al fatto che nel pannello di geni NGS sono già rilevati i geni driver.

Tabella 5.32: Permutation Feature Importance sulle feature genomiche selezionate

Feature	**Importance**	**Std**
ASXL1	0.0790	0.0219
EZH2	0.0496	0.0208
SF3B1	0.0197	0.0169
SRSF2	0.0153	0.0245
U2AF1	0.0139	0.0171
SUZ12	0.0112	0.0095
KMT2A	0.0089	0.0106
CBL	0.0052	0.0122
ZRSR2	0.0000	0.0000
CUX1	0.0000	0.0000
PRPF8	-0.0009	0.0102
NRAS	-0.0036	0.0086
SH2B3	-0.0082	0.0100

È interessante osservare che la mutazione ASXL1 presenta un valore di importance negativo nel modello che integra dati clinici e genomici, mentre assume un'importanza positiva nel modello basato esclusivamente sulle variabili genomiche. Tale comportamento suggerisce che il contributo predittivo di ASXL1 dipenda dal contesto delle altre variabili incluse nel modello. In presenza delle caratteristiche cliniche, gran parte dell'informazione prognostica associata ad ASXL1 risulta già catturata da altri predittori, riducendone il contributo incrementale. Al contrario, nel modello genomico (Tabella 5.32) ASXL1 rappresenta una fonte di

informazione indipendente e la sua permutazione determina una riduzione delle prestazioni del classificatore.

5.7 Analisi di sensibilità

La limitazione dell'analisi ai soli pazienti con grado di fibrosi midollare pari a 0 o 1 determina una lieve riduzione delle prestazioni del modello rispetto all'intera coorte. L'Accuracy diminuisce da 0.891 a 0.873, mentre l'AUC passa da 0.967 a 0.940. Analogamente, il Brier score aumenta da 0.076 a 0.095, indicando una modesta riduzione dell'accuratezza delle probabilità predette.

Tabella 5.33: Confronto delle prestazioni del modello sull'intera coorte e e nei sottogruppi definiti in base al grado di fibrosi.

Metrica	**Intera coorte**	**BOMFibrosi 0–1**
Accuracy	0.891 ± 0.006	0.873 ± 0.058
F1-score	0.896 ± 0.011	0.845 ± 0.076
ROC AUC	0.967 ± 0.011	0.940 ± 0.028
Brier score	0.076 ± 0.008	0.095 ± 0.027

Nonostante tale decremento, tutte le metriche mantengono valori elevati, suggerendo che il modello conserva una buona capacità discriminativa e una soddisfacente calibrazione anche nei pazienti con fibrosi assente o lieve, ovvero nella popolazione clinicamente più difficile da classificare.

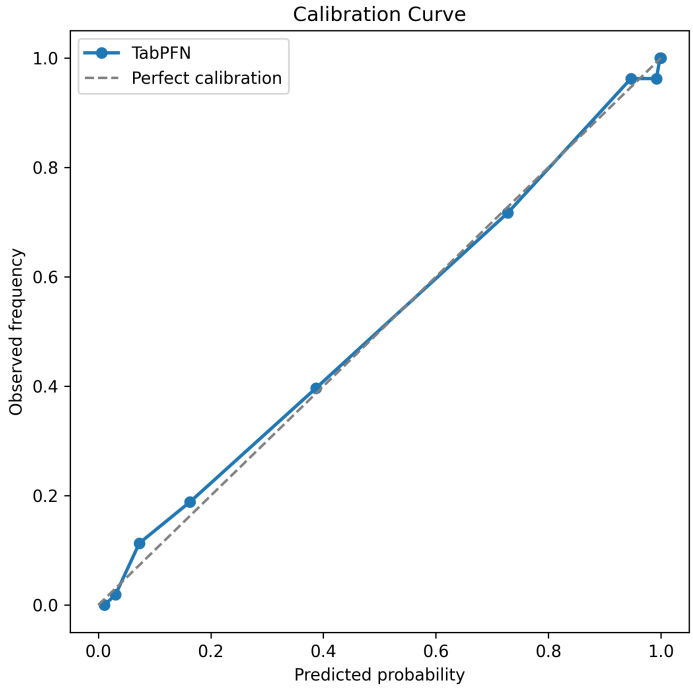

Figura 5.33: Curve di calibrazione del modello su intera coorte.

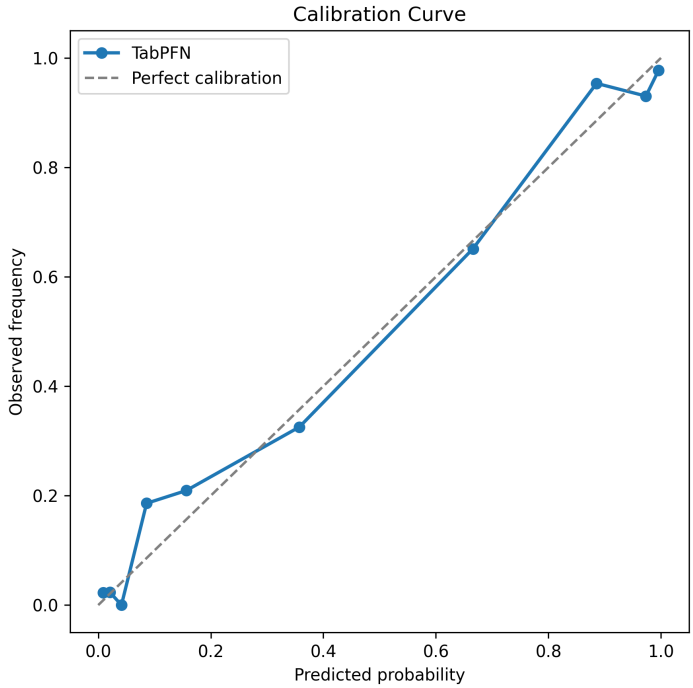

Figura 5.34: Curve di calibrazione del modello su coorte BOMFibrosi 0–1.

5.8 Stratificazione del rischio mediante quartili delle probabilità predette

Come mostrato nella Tabella 5.34, è stato osservato un marcato incremento della prevalenza di fibrosi al crescere del quartile di rischio.

Tabella 5.34: Stratificazione del rischio di fibrosi nel test set sulla base dei quartili delle probabilità predette. E' stato utilizzato il modello TabPFN con strategia *inference*.

Quartile	**Età media**	**Probabilità media**	**Fibrosi (%)**
Q1	51.0	0.082	11.1
Q2	49.2	0.238	29.6
Q3	54.4	0.595	73.1
Q4	59.3	0.977	100.0

Nel primo quartile, caratterizzato da una probabilità media predetta pari a 0.082, la fibrosi era presente nell'11.1% dei pazienti. La prevalenza aumentava al 29.6% nel secondo quartile e al 73.1% nel terzo, raggiungendo il 100% nel quarto quartile, nel quale la probabilità media predetta era pari a 0.977. Tale andamento evidenzia un chiaro gradiente tra rischio predetto e frequenza osservata dell'outcome e suggerisce che il modello sia in grado di identificare, anche in pazienti non utilizzati durante l'addestramento, sottogruppi caratterizzati da un rischio progressivamente crescente di fibrosi.

I risultati in Tabella 5.35 evidenziano un marcato gradiente delle caratteristiche ematologiche attraverso i livelli di rischio identificati dal modello.

La distribuzione della dimensione della milza differisce significativamente tra i quartili ($p < 0.001$). Nei quartili rischio fibrotico più basso la quasi totalità dei pazienti presenta una dimensione splenica pari a zero, mentre nel Q4 tale proporzione si riduce al 19.2%. Nel medesimo quartile, aumenta quindi la frequenza di pazienti con un maggiore coinvolgimento splenico.

Tabella 5.35: Distribuzione delle principali caratteristiche cliniche e laboratoristiche nei quartili di rischio predetto. Per l'età è riportata la mediana; per le variabili categorizzate sono riportati il numero di pazienti e la percentuale tra i soggetti.

Variabile	**Categoria**	**Q1**	**Q2**	**Q3**	**Q4**
Età e sesso					
Età alla diagnosi	Mediana	50.1	48.1	58.7	61.0
Sesso	Maschi	11 (40.7%)	11 (40.7%)	16 (61.5%)	20 (74.1%)
Sesso	Femmine	16 (59.3%)	16 (59.3%)	10 (38.5%)	7 (25.9%)
Milza					
Milza	0	26 (96.3%)	26 (100.0%)	19 (86.4%)	5 (19.2%)
Milza	1–4	1 (3.7%)	0 (0.0%)	1 (4.5%)	14 (53.8%)
Milza	5–9	0 (0.0%)	0 (0.0%)	2 (9.1%)	2 (7.7%)
Milza	≥ 10	0 (0.0%)	0 (0.0%)	0 (0.0%)	5 (19.2%)
Emoglobina					
Emoglobina	< 10	0 (0.0%)	0 (0.0%)	1 (4.0%)	10 (37.0%)
Emoglobina	10–12	0 (0.0%)	1 (3.7%)	4 (16.0%)	7 (25.9%)
Emoglobina	12–14	9 (33.3%)	7 (25.9%)	10 (40.0%)	6 (22.2%)
Emoglobina	≥ 14	18 (66.7%)	19 (70.4%)	10 (40.0%)	4 (14.8%)
Leucociti					
Leucociti	< 5	0 (0.0%)	0 (0.0%)	0 (0.0%)	6 (22.2%)
Leucociti	5–10	17 (63.0%)	14 (51.9%)	15 (60.0%)	9 (33.3%)
Leucociti	10–25	10 (37.0%)	13 (48.1%)	10 (40.0%)	12 (44.4%)
Leucociti	> 25	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Piastrine					
Piastrine	< 100	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (7.4%)
Piastrine	100–450	1 (3.7%)	0 (0.0%)	4 (16.0%)	16 (59.3%)
Piastrine	450–1000	24 (88.9%)	21 (77.8%)	17 (68.0%)	9 (33.3%)
Piastrine	> 1000	2 (7.4%)	6 (22.2%)	4 (16.0%)	0 (0.0%)
LDH					
LDH	< 200	16 (59.3%)	6 (22.2%)	3 (11.5%)	1 (3.7%)
LDH	200–300	11 (40.7%)	20 (74.1%)	13 (50.0%)	9 (33.3%)
LDH	300–400	0 (0.0%)	1 (3.7%)	6 (23.1%)	4 (14.8%)
LDH	> 400	0 (0.0%)	0 (0.0%)	4 (15.4%)	13 (48.1%)
CD34 circolanti					
CD34 circolanti	< 10	27 (100.0%)	26 (96.3%)	24 (92.3%)	14 (51.9%)
CD34 circolanti	10–15	0 (0.0%)	1 (3.7%)	2 (7.7%)	1 (3.7%)
CD34 circolanti	15–100	0 (0.0%)	0 (0.0%)	0 (0.0%)	9 (33.3%)
CD34 circolanti	≥ 100	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (11.1%)

Nota: p-value: età $p = 0.110$; sesso $p = 0.032$; milza $p < 0.001$; emoglobina $p < 0.001$; leucociti $p = 0.002$; piastrine $p < 0.001$; LDH $p < 0.001$; CD34 circolanti $p < 0.001$.

Un andamento particolarmente evidente è osservabile per l'emoglobina ($p < 0.001$). La proporzione di pazienti con emoglobina pari o superiore a 14 diminuisce progressivamente, passando dal 66.7% nel Q1 al 14.8% nel Q4. Parallelamente, nel Q4 il 37.0% dei pazienti presenta valori inferiori a 10 e il 25.9% valori compresi tra 10 e 12, mentre tali categorie sono quasi assenti nei quartili a rischio più basso. Nel complesso, i risultati evidenziano una maggiore frequenza di valori emoglobinici ridotti nei pazienti appartenenti ai gruppi di rischio predetto più elevato.

Anche la distribuzione della conta piastrinica differisce significativamente tra i quartili ($p < 0.001$). Nei gruppi Q1 e Q2 la maggioranza dei pazienti presenta valori compresi tra 450 e 1000, rispettivamente nell'88.9% e nel 77.8% dei casi. Nel Q4, invece, questa proporzione si riduce al 33.3%, mentre aumenta considerevolmente la frequenza di valori compresi tra 100 e 450, pari al 59.3%. Nel Q4 è inoltre presente il 7.4% dei pazienti con conta inferiore a 100. Tale distribuzione evidenzia una maggiore frequenza di valori piastrinici ridotti nel gruppo caratterizzato dal rischio predetto più elevato.

Per quanto riguarda l'LDH, è stata osservata una differenza significativa nella distribuzione delle categorie tra i quattro quartili ($p < 0.001$). La proporzione di pazienti con valori inferiori a 200 diminuisce dal 59.3% nel Q1 al 3.7% nel Q4, mentre nel Q4 il 48.1% dei pazienti presenta valori superiori a 400. Anche in questo caso, pertanto, l'aumento del rischio predetto è associato a una maggiore frequenza di valori elevati della variabile.

La distribuzione dei CD34 circolanti mostra una differenza significativa tra i gruppi ($p < 0.001$). Nei quartili Q1 e Q2 la quasi totalità dei pazienti presenta valori inferiori a 10, mentre nel Q4 tale proporzione scende al 51.9%. Nel medesimo quartile, il 33.3% dei pazienti presenta valori compresi tra 15 e 100 e l'11.1% valori pari o superiori a 100. Questo risultato evidenzia una maggiore presenza di valori elevati di CD34 circolanti tra i pazienti caratterizzati da probabilità predette di fibrosi più elevate. Considerata l'elevata frequenza di dati mancanti per questa variabile, tale risultato deve tuttavia essere interpretato con cautela e riferito esclusivamente ai pazienti con valore disponibile.

Per la conta leucocitaria è stata osservata una differenza statisticamente significativa nella distribuzione delle categorie ($p = 0.002$). Tale differenza è principalmente riconducibile alla presenza, nel Q4, di una maggiore proporzione di pazienti con valori inferiori a 5, pari al 22.2%, categoria assente negli altri quartili. Non emerge tuttavia un andamento monotono altrettanto evidente rispetto a quello osservato per emoglobina, piastrine, LDH e dimensione splenica.

Infine, la distribuzione del sesso differisce significativamente tra i quartili ($p = 0.032$), con una progressiva maggiore rappresentazione dei pazienti di sesso maschile nei gruppi a rischio più elevato: la proporzione di maschi passa infatti dal 40.7% nei quartili Q1 e Q2 al 61.5% nel Q3 e al 74.1% nel Q4. Al contrario, l'età alla diagnosi non mostra differenze statisticamente significative tra i quattro gruppi ($p = 0.110$), nonostante la mediana aumenti da 50.1 anni nel Q1 a 61.0 anni nel Q4.

Nel complesso, la stratificazione evidenzia che l'aumento della probabilità predetta dal modello si accompagna a una diversa distribuzione di numerose caratteristiche cliniche e laboratoristiche. Infatti, i pazienti appartenenti ai quartili di rischio più elevato mostrano più frequentemente coinvolgimento splenico, valori ridotti di emoglobina e piastrine e valori elevati di LDH e CD34 circolanti. Questi risultati sono coerenti con la presenza di un gradiente clinico tra i livelli di rischio identificati dal modello e supportano l'utilità della stratificazione mediante probabilità predette per caratterizzare sottogruppi di pazienti con profili clinici differenti.

La Tabella 5.36 mostra le prestazioni del modello su pazienti caratterizzati da una probabilità predetta compresa tra 0,40 e 0,60, identificati come popolazione borderline rispetto alla soglia decisionale di 0,50. Nel test set, tale fascia comprendeva 11 pazienti su 107 (10,3%). Tra questi, 5 pazienti (45,5%) sono stati classificati correttamente, mentre 6 (54,5%) sono risultati erroneamente classificati. Al contrario, tra i 96 pazienti al di fuori della zona grigia, 82 (85,4%) sono stati classificati correttamente e 14 (14,6%) erroneamente.

Il tasso di errore è risultato pertanto sostanzialmente maggiore nella zona grigia rispetto al resto della coorte di test (54,5% vs 14,6%). Il confronto mediante test esatto di Fisher ha evidenziato un'associazione statisticamente significativa tra appartenenza alla zona grigia ed errore di classificazione (OR = 7,03; p = 0,005).

Tabella 5.36: Prestazioni del modello nei pazienti appartenenti alla zona grigia di probabilità predetta (0,40–0,60) rispetto ai pazienti al di fuori di tale intervallo.

Metrica	**Zona grigia**	**Fuori zona grigia**
N pazienti	11	96
Pazienti corretti	5 (45,5%)	82 (85,4%)
Pazienti errati	6 (54,5%)	14 (14,6%)
Accuracy	0,455	0,854
Error rate	0,545	0,146
Test esatto di Fisher: OR = 7,03; p = 0,005		

Inoltre, tra gli 11 pazienti della zona grigia, 8 sono realmente fibrotici e 3 realmente non fibrotici. Dei fibrotici, 5 sono stati classificati come non fibrotici.

Tabella 5.37: Matrice di confusione dei pazienti appartenenti alla zona grigia di probabilità predetta (0,40–0,60).

	Classe predetta	
Classe reale	**Non fibrotica**	**Fibrotica**
Non fibrotica	2 (TN)	1 (FP)
Fibrotica	5 (FN)	3 (TP)

Considerata la limitata numerosità della popolazione borderline (n=11), è stata effettuata un'analisi esclusivamente descrittiva delle principali caratteristiche cliniche, che nel complesso non mostrava un profilo clinico univoco, ma comprendeva pazienti con caratteristiche eterogenee.

Capitolo 6

Discussione

In questo capitolo verranno discussi i principali risultati ottenuti nel corso dello studio. In particolare, il contributo delle diverse fonti informative alla predizione della diagnosi fibrotica: dalla capacità predittiva dei dati genomici derivati da NGS al possibile beneficio clinico derivante dall'esclusione delle informazioni relative alla biopsia ossea. Infine, i limiti e considerazioni metodologiche appartenenti alla costruzione degli embedding e dei Tabular Foundation Models studiati.

6.1 Capacità predittiva dei dati NGS

I risultati riportati nella Tabella 5.13 forniscono una prima valutazione della capacità predittiva dell'informazione genomica, utilizzata in assenza di variabili cliniche. Nel complesso, le prestazioni ottenute indicano che i dati NGS contengono un'informazione predittiva non trascurabile rispetto alla classificazione del fenotipo fibrotico, ma che tale informazione risulta tuttavia limitata e non sufficiente a garantire una discriminazione ottimale tra le due classi.

È rilevante osservare che la classe fibrotica presenta una precisione relativamente elevata (0.84). Ciò indica che l'informazione genomica permette, quando il modello identifica un campione come fibrotico, di effettuare una predizione generalmente corretta; la difficoltà

principale riguarda piuttosto la capacità di riconoscere la totalità dei soggetti fibrotici. Considerando inoltre che le numerosità delle due classi sono comparabili (50 campioni non fibrotici e 57 fibrotici), la differenza osservata nelle prestazioni non appare riconducibile a un marcato sbilanciamento delle classi.

Nel complesso, questi risultati suggeriscono quindi che i dati NGS possiedono un potere predittivo parziale rispetto al fenotipo di fibrosi, ma che la sola informazione genomica non sembra essere sufficiente a descrivere completamente la variabilità fenotipica osservata.
Tuttavia, la presenza di un'informazione predittiva nei dati genomici potrebbe rappresentare un possibile punto di partenza per lo sviluppo di modelli capaci di fornire elementi utili per predire lo sviluppo di fibrosi successiva.

6.2 Valutazione del rapporto beneficio clinico–prestazionale

La Permutation Feature Importance ha individuato il grado di fibrosi midollare come la variabile più importante, che da sola consente di raggiungere già un valore di F1 elevato. Esso è, infatti, uno degli elementi utilizzati nella pratica clinica per distinguere queste condizioni; pertanto, il risultato è del tutto atteso.

Essendo la biopsia ossea parte del processo diagnostico, è stato interessante studiare le capacità predittive dei modelli usando solo dati disponibili prima della biopsia (esami del sangue, mutazioni, dati clinici iniziali). I risultati delle performance dei modelli (Tabella 5.22 e Tabella 5.23) hanno mostrato una lieve riduzione delle capacità predittive, che dovrebbe essere interpretata alla luce del rapporto tra accuratezza predittiva e beneficio clinico. Una modesta perdita di performance potrebbe infatti risultare accettabile qualora sia compensata dalla possibilità di ridurre il ricorso a procedure invasive, migliorando l'accettabilità del modello per il paziente e ampliandone l'applicabilità in contesti clinici nei quali la biopsia non sia indicata, non sia disponibile o non possa essere eseguita.

È importante sottolineare però, che i risultati ottenuti dimostrano esclusivamente che la variabile relativa alla biopsia potrebbe non essere necessaria come input del modello predittivo

e non implicano, di per sé, che la biopsia possa essere omessa nel percorso diagnostico del paziente.

6.3 Limiti e considerazioni metodologiche

I risultati riportati nelle Tabelle 5.18 e 5.19 suggeriscono che le rappresentazioni apprese da `Mut2Vec` non aggiungono informazione discriminativa utile per `TabICL`.

Una possibile spiegazione è che il modello sia già in grado di estrarre efficacemente le relazioni presenti nelle feature originali e che gli embedding introducano una rappresentazione ridondante o meno allineata con il meccanismo di apprendimento del modello. Inoltre, gli embedding potrebbero modificare la distribuzione delle feature, rendendo meno efficace la capacità di `TabICL` di sfruttare i pattern presenti nei dati tabellari originali.

In più, gli embedding `Mut2Vec` non sono stati addestrati specificamente su neoplasie mieloproliferative, quindi il contesto di co-occorrenza potrebbe essere diverso. Si potrebbe successivamente reimpostare il modello Skip-Gram solo su pazienti affetti da MPN, per aggiornare gli embedding in un contesto MPN-specifico. Inoltre, sebbene una semplice media degli embedding (mean pooling) rappresenti una baseline robusta e interpretabile, con il mean pooling tutte le mutazioni pesano uguale.

Nel miglior modello osservato, TabPFN con embedding, l'assenza di differenze apprezzabili fra le diverse strategie di tuning nelle diverse configurazioni di parametri (Tabella 5.24), può essere attribuita a diversi fattori: da un lato, il modello potrebbe essere già sufficientemente robusto da convergere verso una soluzione analoga per tutte le combinazioni esplorate; dall'altro, il limite prestazionale può essere determinato principalmente dalle caratteristiche del dataset piuttosto che dalla scelta degli iperparametri.

A parità di prestazioni, si privilegia la strategia con il minor costo computazionale.

Capitolo 7

Conclusioni

Gli obiettivi della tesi erano valutare le prestazioni dei recenti Tabular Foundation Models nella predizione del fenotipo fibrotico in pazienti affetti da neoplasie mieloproliferative e studiare rappresentazioni vettoriali aggiuntive dei dati genomici, sotto forma di embedding, con lo scopo di verificare se una trasformazione più informativa delle variabili potesse compensare, almeno in parte, la sparsità del dataset.

Sebbene gli embedding abbiano fornito una rappresentazione più compatta e densa dei dati, la loro integrazione ha avuto un impatto limitato sulle prestazioni, già ottime, dei modelli.

Tra i modelli valutati, TabPFN in configurazione `inference` con l'integrazione degli embedding `Mut2Vec` ha ottenuto le performance complessivamente migliori. La configurazione `inference` ha ottenuto le migliori prestazioni senza richiedere un riaddestramento completo del modello sul dataset considerato. Questo risultato è particolarmente rilevante in presenza di una coorte di dimensioni limitate, poiché riduce il rischio di sovradattamento e sfrutta le capacità di generalizzazione apprese durante il pre-training.

Le prestazioni ottenute da TabPFN sono in linea con la motivazione alla base del modello, sviluppato specificamente per affrontare problemi di classificazione su dataset tabulari di dimensioni ridotte o moderate. TabPFN utilizza un approccio di prior-data fitted inference: il

modello viene addestrato offline su dataset sintetici generati da un prior e, successivamente, applica una forma di apprendimento in-context al nuovo dataset senza richiedere un'ottimizzazione estesa degli iperparametri. Nel lavoro originale, il modello aveva mostrato risultati competitivi o superiori rispetto a metodi tradizionali su numerosi benchmark tabulari [33, 8].

Per concludere, un possibile sviluppo futuro del presente studio potrebbe consistere nell'estensione dell'analisi a una dimensione longitudinale, includendo dati acquisiti in differenti momenti del percorso clinico del paziente. Ciò permetterebbe di valutare l'evoluzione temporale delle variabili diagnostiche considerate e di identificare variazioni associate alla progressione della malattia o al miglioramento delle condizioni cliniche.

L'integrazione delle informazioni relative ai trattamenti terapeutici somministrati consentirebbe di studiare come le caratteristiche osservate al basale e la loro evoluzione nel tempo siano correlate al trattamento.

Pertanto, la disponibilità di dati longitudinali e di trattamento renderebbe possibile lo sviluppo di modelli prognostici finalizzati alla previsione di outcome clinicamente rilevanti, quali la progressione della malattia, la sopravvivenza libera da progressione o la sopravvivenza globale. Tali modelli potrebbero contribuire a una migliore stratificazione del rischio dei pazienti e supportare approcci di medicina personalizzata, favorendo decisioni terapeutiche più mirate e tempestive.

Bibliografia

[1] Elena Vuelta, Ignacio García-Tuñón, Patricia Hernández-Carabias, Lucía Méndez, and Manuel Sánchez-Martín. Future approaches for treating chronic myeloid leukemia: Crispr therapy. *Biology*, 10(2):118, 2021. DOI: 10.3390/biology10020118.

[2] Donal P McLornan, Anna L Godfrey, Anna Green, Rebecca Frewin, Siamak Arami, Jessica Brady, Nauman M Butt, Catherine Cargo, Joanne Ewing, Sebastian Francis, et al. Diagnosis and evaluation of prognosis of myelofibrosis: A british society for haematology guideline. 2023. DOI: 10.1111/bjh.19164.

[3] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019. DOI: 10.1056/NEJMra1814259.

[4] Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial intelligence review*, 56(9):10345–10425, 2023. DOI: 10.1007/s10462-023-10419-1.

[5] Francesca Incitti, Federico Urli, and Lauro Snidaro. Beyond word embeddings: A survey. *Information Fusion*, 89:418–436, 2023. DOI: 10.1016/j.inffus.2022.08.024.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. DOI: 10.48550/arXiv.1706.03762.

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al.

On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. DOI: 10.48550/arXiv.2108.07258.

[8] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. DOI: 10.1038/s41586-024-08328-6.

[9] Joseph D Khoury, Eric Solary, Oussama Abla, Yassmine Akkari, Rita Alaggio, Jane F Apperley, Rafael Bejar, Emilio Berti, Lambert Busque, John KC Chan, et al. The 5th edition of the world health organization classification of haematolymphoid tumours: myeloid and histiocytic/dendritic neoplasms. *Leukemia*, 36(7):1703–1719, 2022. DOI: 10.1038/s41375-022-01613-1.

[10] Jürgen Thiele, Hans Michael Kvasnicka, Attilio Orazi, Umberto Gianelli, Naseema Gangat, Alessandro M Vannucchi, Tiziano Barbui, Daniel A Arber, and Ayalew Tefferi. The international consensus classification of myeloid neoplasms and acute leukemias: myeloproliferative neoplasms. *American journal of hematology*, 98(1):166–179, 2023. DOI: 10.1002/ajh.26751.

[11] Sachi Singhal, Rahul Mishra, Sankalp Arora, Noha Sharafeldin, Ruchi Desai, Tania Jain, and Pankit Vachhani. Incidence, prevalence, and survival outcomes of patients with myeloproliferative neoplasms in the united states: a seer database analysis, years 2000–2021. *Research Square*, pages rs–3, 2026. DOI: 10.1038/s41375-026-03124-9.

[12] Judith Staerk and Stefan N Constantinescu. The jak-stat pathway and hematopoietic stem cells from the jak2 v617f perspective. *Jak-Stat*, 1(3):184–190, 2012. DOI: 10.4161/jkst.22071.

[13] Michael Stephan Bader and Sara Christina Meyer. Jak2 in myeloproliferative neoplasms: still a protagonist. *Pharmaceuticals*, 15(2):160, 2022. DOI: 10.3390/ph15020160.

[14] Alessandro Maria Vannucchi, TL Lasho, Paola Guglielmelli, Flavia Biamonte, A Pardanani, A Pereira, C Finke, J Score, N Gangat, Carmela Mannarelli, et al. Mutations and prognosis in primary myelofibrosis. *Leukemia*, 27(9):1861–1869, 2013. DOI: 10.1038/leu.2013.119.

[15] Damien Luque Paz, Jeremie Riou, Emmanuelle Verger, Bruno Cassinat, Aurélie Chauveau, Jean-Christophe Ianotto, Brigitte Dupriez, Françoise Boyer, Maxime Renard, Olivier Mansier, et al. Genomic analysis of primary and secondary myelofibrosis redefines the prognostic impact of asxl1 mutations: a fim study. *Blood Advances*, 5(5):1442–1451, 2021. DOI: 10.1182/bloodadvances.2020003444.

[16] Ross L Levine, Animesh Pardanani, Ayalew Tefferi, and D Gary Gilliland. Role of jak2 in the pathogenesis and therapy of myeloproliferative disorders. *Nature reviews cancer*, 7(9):673–683, 2007. DOI: 10.1038/nrc2210.

[17] Elisa Rumi, Daniela Pietra, Virginia Ferretti, Thorsten Klampfl, Ashot S Harutyunyan, Jelena D Milosevic, Nicole CC Them, Tiina Berg, Chiara Elena, Ilaria C Casetti, et al. Jak2 or calr mutation status defines subtypes of essential thrombocythemia with substantially different clinical course and outcomes. *Blood, The Journal of the American Society of Hematology*, 123(10):1544–1551, 2014. DOI: 10.1182/blood-2013-11-539098.

[18] Ayalew Tefferi, Alessandro Maria Vannucchi, Tiziano Barbui, et al. Essential thrombocythemia: 2024 update on diagnosis, risk stratification, and management. *American journal of hematology*, 99:697–718, 2024. DOI: 10.1002/ajh.27216.

[19] William Vainchenker and Robert Kralovics. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. *Blood, The Journal of the American Society of Hematology*, 129(6):667–679, 2017. DOI: 10.1182/blood-2016-10-695940.

[20] Sabrina Fowlkes, Cindy Murray, Adrienne Fulford, Tammy De Gelder, and Nancy Siddiq. Myeloproliferative neoplasms (mpns)–part 1: An overview of the diagnosis and treatment

of the "classical" mpns. *Canadian Oncology Nursing Journal*, 28(4):262, 2018. DOI: 10.5737/23688076284262268.

[21] Aziz Nazha, Joseph D Khoury, Raajit K Rampal, and Naval Daver. Fibrogenesis in primary myelofibrosis: diagnostic, clinical, and therapeutic implications. *The oncologist*, 20(10):1154–1160, 2015. DOI: 10.1634/theoncologist.2015-0094.

[22] Gaurav Mendiratta, Eugene Ke, Meraj Aziz, David Liarakos, Melinda Tong, and Edward C Stites. Cancer gene mutation frequencies for the us population. *Nature communications*, 12(1):5961, 2021. DOI: 10.1038/s41467-021-26213-y.

[23] Ronald A Fisher. The logic of inductive inference. *Journal of the royal statistical society*, 98(1):39–82, 1935. DOI: 10.2307/2342435.

[24] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[25] Aasim Ayaz Wani. Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions. *PeerJ Computer Science*, 11:e3025, 2025. DOI: 10.7717/peerj-cs.3025.

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. DOI: 10.48550/arXiv.1301.3781.

[27] Sunkyu Kim, Heewon Lee, Keonwoo Kim, and Jaewoo Kang. Mut2vec: distributed representation of cancerous mutations. *BMC medical genomics*, 11(Suppl 2):33, 2018. DOI: 10.1186/s12920-018-0349-7.

[28] International Cancer Genome Consortium et al. International network of cancer genome projects. *Nature*, 464(7291):993, 2010. DOI: 10.1038/nature08987.

[29] David Oniani, Chen Wang, Yiqing Zhao, Andrew Wen, Hongfang Liu, and Feichen Shen. Leveraging a joint of phenotypic and genetic features on cancer patient subgrouping. *arXiv preprint arXiv:2103.16316*, 2021. DOI: 10.48550/arXiv.2103.16316.

[30] Prima Sanjaya, Katri Maljanen, Riku Katainen, Sebastian M Waszak, Lauri A Aaltonen, Oliver Stegle, Jan O Korbel, and Esa Pitkänen. Mutation-attention (muat): deep representation learning of somatic mutations for tumour typing and subtyping. *Genome medicine*, 15(1):47, 2023. DOI: 10.1186/s13073-023-01204-4.

[31] Aditya Tanna, Pratinav Seth, Mohamed Bouadi, and Vinay Kumar Sankarapu. Tabtune: A unified library for inference and fine-tuning tabular foundation models. In *Companion Proceedings of the ACM Web Conference 2026*, pages 204–207, 2026. DOI: 10.1145/3774905.3793137.

[32] Parul Pandey. Exploring tabpfn: A foundation model built for tabular data. https://pandeyparul.medium.com/exploring-tabpfn-a-foundation-model-built-for-tabular-data-3ad3177cc17f, jan 2026. Medium.

[33] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022. DOI: 10.48550/arXiv.2207.01848.

[34] Jingang Qu, David HolzmÃžller, GaÃl Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. *arXiv preprint arXiv:2502.05564*, 2025. DOI: 10.5555/3780338.3782366.

[35] Xiyuan Zhang, Danielle Maddix Robinson, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael Mahoney, et al. Mitra: Mixed synthetic priors for enhancing tabular foundation models. *Advances in neural information processing systems*, 38:15795–15840, 2026. DOI: 10.48550/arXiv.2510.21204.

[36] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. DOI: 10.48550/arXiv.2005.14165.

[37] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. DOI: 10.48550/arXiv.1902.00751.

[38] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. DOI: 10.48550/arXiv.2106.09685.

[39] Alfonso Monaco, Ester Pantaleo, Nicola Amoroso, Antonio Lacalamita, Claudio Lo Giudice, Adriano Fonzino, Bruno Fosso, Ernesto Picardi, Sabina Tangaro, Graziano Pesole, et al. A primer on machine learning techniques for genomic applications. 2021. DOI: 10.1016/j.csbj.2021.07.021.

[40] Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012. DOI: 10.1016/j.ygeno.2012.04.003.

[41] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018. DOI: 10.48550/arXiv.1706.09516.

[42] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010. DOI: 10.1093/bioinformatics/btq134.

[43] Christoph Molnar. The interpretability tax on tabular foundation models. https://mindfulmodeler.substack.com/p/tabular-foundation-models-break-the, mar 2026. Mindful Modeler.

[44] Ana Carolina Alba, Thomas Agoritsas, Michael Walsh, Steven Hanna, Alfonso Iorio, PJ Devereaux, Thomas McGinn, and Gordon Guyatt. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. *Jama*, 318(14): 1377–1384, 2017. DOI: 10.1001/jama.2017.12126.

Appendice A

Risultati Wilcoxon signed-rank test

Tabella A.1: Risultati sulla 10-fold-cv del test di Wilcoxon signed-rank per il confronto tra i modelli addestrati con e senza embedding, utilizzando l'F1-score come metrica di valutazione. I p-value sono stati corretti mediante il metodo di Holm. E' stato adottato un livello di significatività pari a $\alpha = 0.05$.

Modello	**Strategia**	**Statistica**	**p-value**	**p-value corretto**
Mitra	inference	8	0.3750	0.6563
Mitra	peft	3	0.0195	0.0977
TabICL	base-ft	1	0.0039	0.0234
TabICL	inference	0	0.0020	0.0137
TabICL	peft	10	0.0840	0.3359
TabPFN	base-ft	6	0.2188	0.6563
TabPFN	inference	6	0.2188	0.6563

Tabella A.2: Risultati sulla 10-fold-cv del test di Wilcoxon signed-rank per il confronto tra i modelli addestrati con e senza embedding, utilizzando la ROC AUC come metrica di valutazione. I p-value sono stati corretti mediante il metodo di Holm.

Modello	**Strategia**	**Statistica**	**p-value**	**p-value corretto**
Mitra	inference	12	0.1309	0.3926
Mitra	peft	3	0.0098	0.0391
TabICL	base-ft	0	0.0020	0.0137
TabICL	inference	0	0.0020	0.0137
TabICL	peft	0	0.0039	0.0195
TabPFN	base-ft	25	0.8457	1.0000
TabPFN	inference	25	0.8457	1.0000

Appendice B

Forward Stepwise Feature Selection

```
def forward_stepwise_tabpfn(
    X,
    y,
    k=5,
    metric="f1_score"
):

    best_global_score = -np.inf
    min_improvement = 0.002
    patience = 2
    no_improvement = 0

    remaining = list(X.columns)
    selected = []

    history = []

    skf = StratifiedKFold(
        n_splits=k,
```

```
        shuffle=True,
        random_state=42
    )

    while len(remaining) > 0:

        candidate_scores = []

        for feature in remaining:

            current_features = selected + [feature]
            fold_scores = []

            for train_idx, val_idx in skf.split(X, y):

                X_train_fold = X.iloc[train_idx][current_features]
                X_val_fold = X.iloc[val_idx][current_features]

                y_train_fold = y.iloc[train_idx]
                y_val_fold = y.iloc[val_idx]

                pipeline = TabularPipeline(
                    model_name="TabPFN",
                    tuning_strategy="inference"
                )

                try:
                  pipeline.fit(X_train_fold, y_train_fold)
                  metrics = pipeline.evaluate(
                    X_val_fold,
```

```
50                      y_val_fold
51                  )
52                  fold_scores.append(metrics[metric])
53
54              except ValueError:
55                  # la feature    costante in almeno un fold
56                  fold_scores = []
57                  break
58
59          if len(fold_scores) == 0:
60              continue
61
62          mean_score = np.mean(fold_scores)
63          candidate_scores.append(
64              (feature, mean_score)
65          )
66
67      best_feature, best_score = max(
68          candidate_scores,
69          key=lambda x: x[1]
70      )
71
72      improvement = best_score - best_global_score
73      selected.append(best_feature)
74      remaining.remove(best_feature)
75
76      history.append({
77          "n_features": len(selected),
78          "added_feature": best_feature,
79          "mean_f1": best_score,
```

```
80              "improvement": improvement,
81              "selected_features": selected.copy()
82          })
83
84          # Controllo del miglioramento
85          if improvement >= min_improvement:
86            best_global_score = best_score
87            no_improvement = 0
88          else:
89            no_improvement += 1
90
91          # Early stopping
92          if no_improvement >= patience:
93              print("\nSTOP")
94              break
95
96      return pd.DataFrame(history)
```

Appendice C

Decision Framework per la selezione dei Tabular Foundation Models

Il Decision Framework [31] utilizzato per la selezione dei modelli è organizzato secondo tre criteri principali: dimensione del dataset, tipologia delle feature e risorse computazionali.

C.1 By Dataset Size

- **Dataset Size $<$ 10K righe**
  - **Raccomandato:** TabPFN, Mitra
  - **Alternativa:** TabICL con valore di `n_estimators` piccolo
- **Dataset Size: 10K–100K righe**
  - **Raccomandato:** TabICL
  - **Alternativa:** Mitra, OrionMSP, OrionBix
- **Dataset Size: 100K–1M righe**
  - **Reccomandato:** OrionMSP, OrionBix, TabDPT
  - **Alternativa:** TabICL con valore di `n_estimators` grande
- **Dataset Size $>$ 1M rows**

  - **Reccomandato:** TabDPT
  - **Alternativa:** OrionBix con chunked training

C.2 By Feature Types

- **Primarily Numerical**
  - **Best:** TabDPT, TabICL
  - **Reason:** Pipeline efficienti di scaling e normalizzazione
- **Primarily Categorical**
  - **Best:** TabPFN, ContextTab
  - **Reason:** Codifica categorica specializzata
- **Mixed (Numerical + Categorical)**
  - **Best:** TabICL, OrionMSP, OrionBix, Mitra
  - **Reason:** Gestione equilibrata di entrambe le tipologie
- **Text/Semantic Features**
  - **Best:** ContextTab
  - **Reason:** supporta built-in text embedding

C.3 By Computational Budget

- **Limited Resources (<8GB GPU)**
  - **Raccomandato:** TabPFN (inference), TabICL (PEFT)
  - **Strategy:** usa `peft` tuning strategy
- **Moderate Resources (8–16GB GPU)**
  - **Raccomandato:** TabICL, OrionMSP, OrionBix
  - **Strategy:** `base-ft` o `peft`
- **Ample Resources (>16GB GPU)**
  - **Raccomandato:** TabDPT, Mitra, OrionBix
  - **Strategy:** `base-ft` with mixed precision

Appendice D

Elenco delle features cliniche

Tabella D.1: Variabili cliniche considerate nell'analisi e relativa tipologia.

Feature	**Tipo di variabile**
Data di nascita	Temporale
Sesso	Categorica
Conta dei leucociti	Numerica
Concentrazione di emoglobina	Numerica
Valore dell'ematocrito	Numerica
Conta piastrinica	Numerica
Livello di lattato deidrogenasi (LDH)	Numerica
Dimensioni della milza (cm)	Numerica
Conta dei CD34 circolanti	Numerica
Livelli di eritropoietina	Numerica
Presenza di biopsia ossea	Binaria
Percentuale di cellularità midollare	Numerica
Adeguatezza della cellularità midollare rispetto all'età del paziente	Numerica
Grado di fibrosi midollare	Ordinale
Percentuale di mutazione allelica JAK2V617F	Numerica