UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE
CORSO DI LAUREA MAGISTRALE IN INFORMATICA
ANNO ACCADEMICO 2023/2024

Integrazione di dati multimodali tramite metodi basati su reti di pazienti

Luigi Santise

Relatore: Elena Casiraghi, Correlatore: Jessica Gliozzo

Contesto e motivazione

- Acquisizione di grandi quantità di dati multimodali
- Descrizioni a diversi livelli biomolecolari dei pazienti
- Avanzamento della medicina di precisione

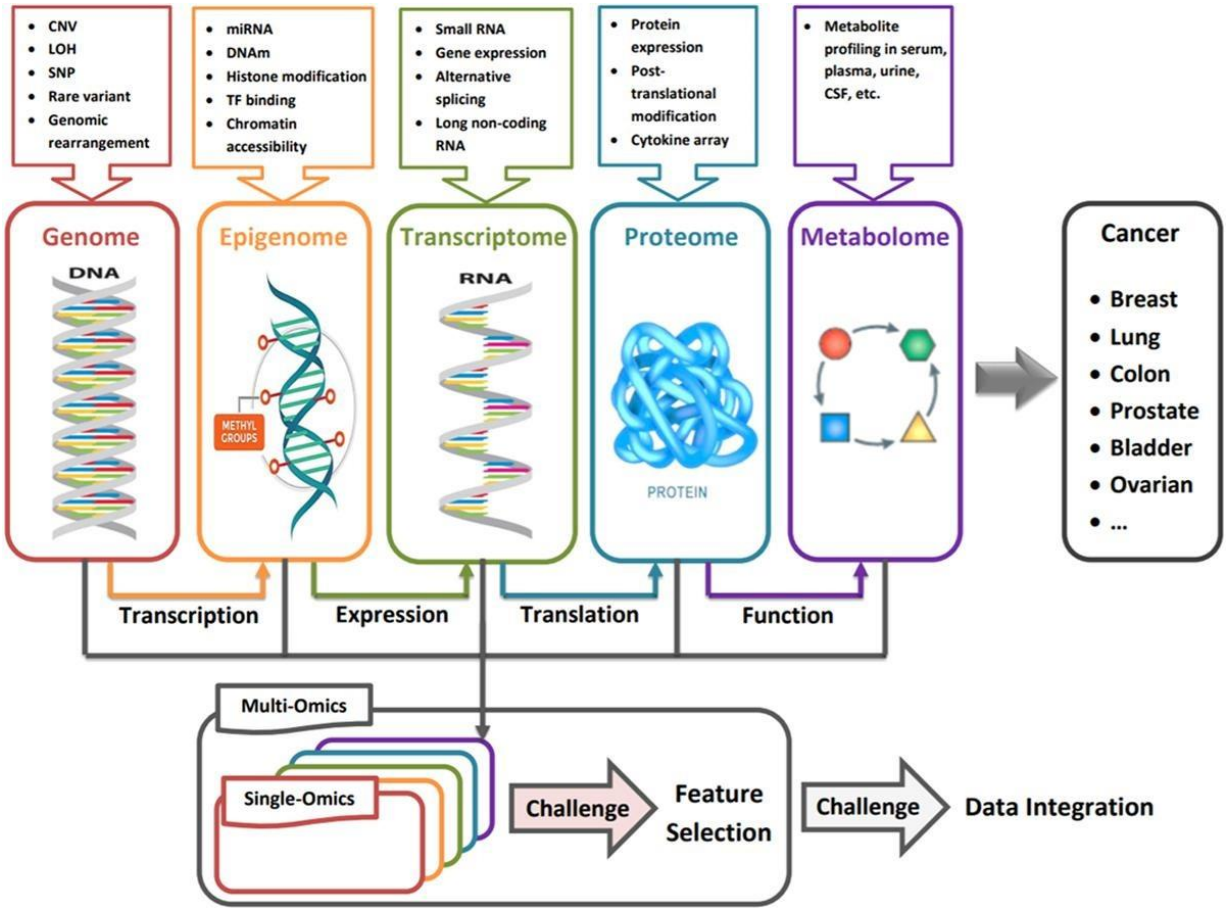


Immagine da Momeni et al., J. Biomed. Inform., 2020.

Metodologia – Costruzione delle reti di similarità I

- Reti di similarità tra pazienti *(Pai et al., J. Mol. Biol., 2018)*
  - Nodi: pazienti
  - Archi: similarità tra i pazienti
- Integrazione con Similarity Network Fusion (SNF) *(Wang et al., Nat. Methods, 2014)*
- K-Nearest Neighbors (KNN) *(Guo et al., CoopIS, 2003)*
  - Sparsificazione della matrice di affinità

Metodologia – Costruzione delle reti di similarità II

- SNF: message passing per la diffusione di informazioni tra reti di similarità unimodali

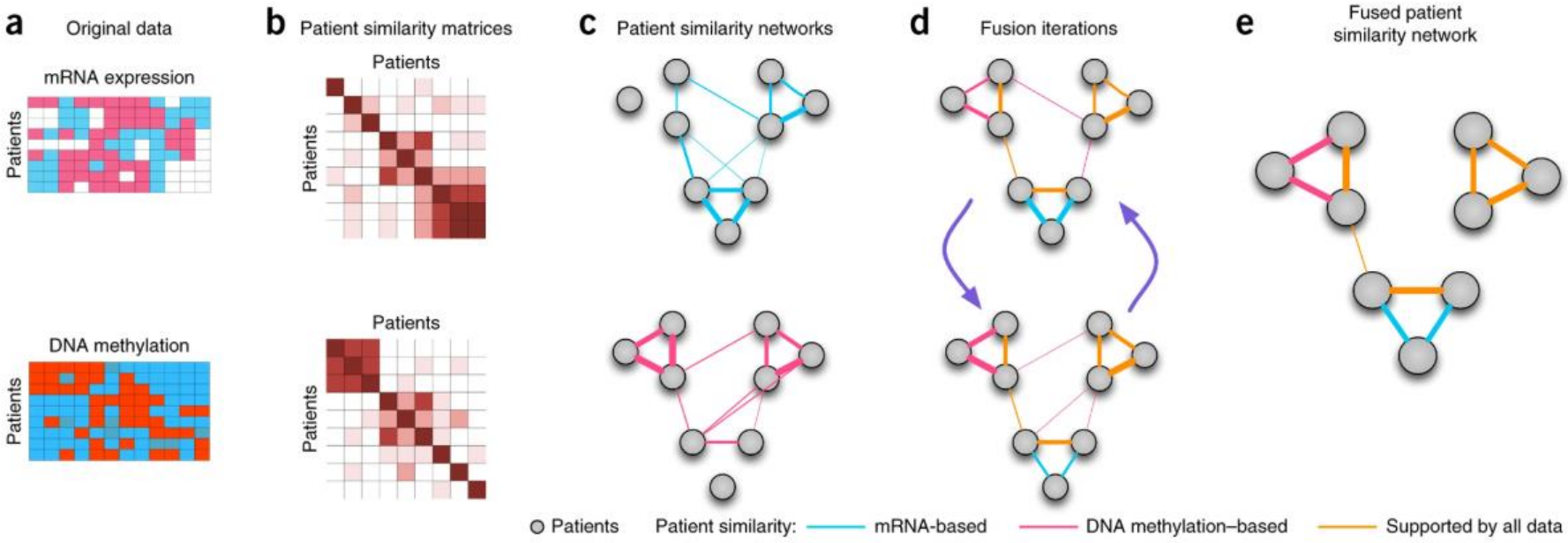
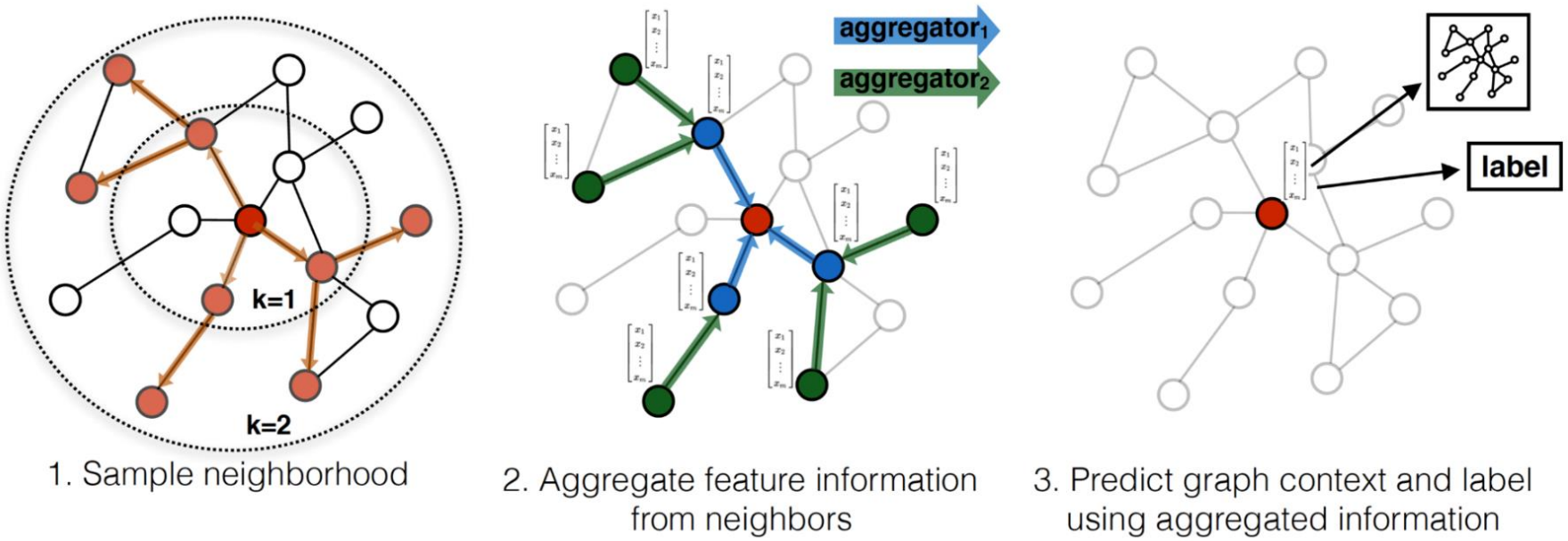


Immagine da Wang et al., Nat. Methods, 2014.

Metodologia - Modello di apprendimento

- GraphSAGE: modello induttivo che permette di generare embedding per nodi mai visti prima *(Hamilton et al., NIPS, 2017)*
  - Graph Neural Network (GNN): apprende direttamente dalla struttura del grafo *(Scarselli et al., IEEE, 2008)*
  - Aggregatore Mean



1. Sample neighborhood

2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

Immagine da Hamilton et al., NIPS, 2017.

Dataset utilizzato

- TCGA-BRCA (The Cancer Genome Atlas Breast Cancer): dati di pazienti affetti da tumore al seno *(Tomczak et al., TCGA, 2014)*
  - Metodo PAM50: Basal, HER2, Luminal A, Luminal B e Normal *(Bastien et al., BMC, 2012)*

- Modalità omiche considerate: mRNA, miRNA e DNAm.

Esperimenti condotti

- Sono stati mantenuti solo i campioni comuni tra le tre modalità omiche (696 pazienti)

- Primo esperimento: SNF + KNN
- Secondo esperimento: KNN + SNF + KNN

Utilizzo delle liste delle feature fornite da MOGDx

- Terzo esperimento: KNN + SNF + KNN
  - Per rimuovere il bias introdotto dalla selezione supervisionata effettuata utilizzando l'intero dataset

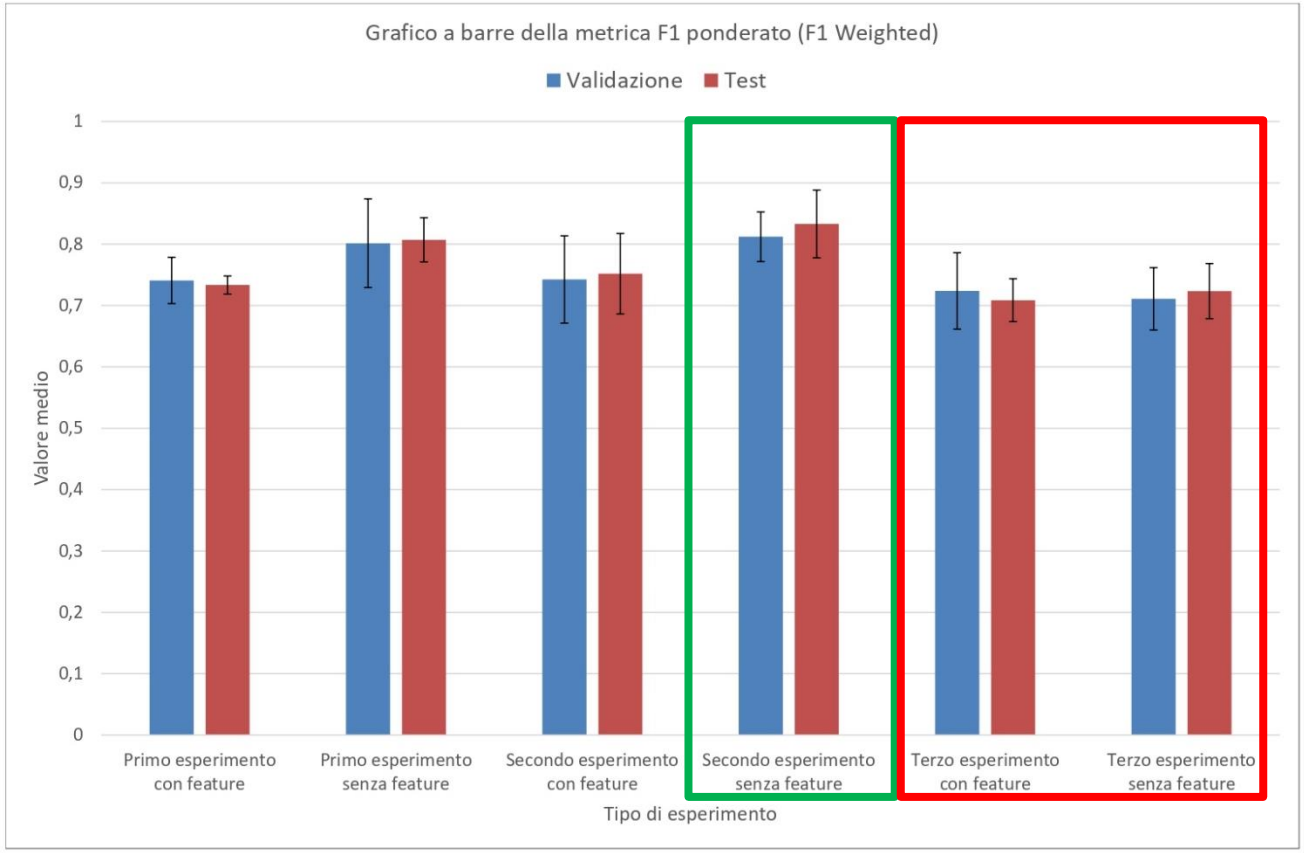
Selezione delle 500 feature con la maggiore variabilità

Addestramento del modello

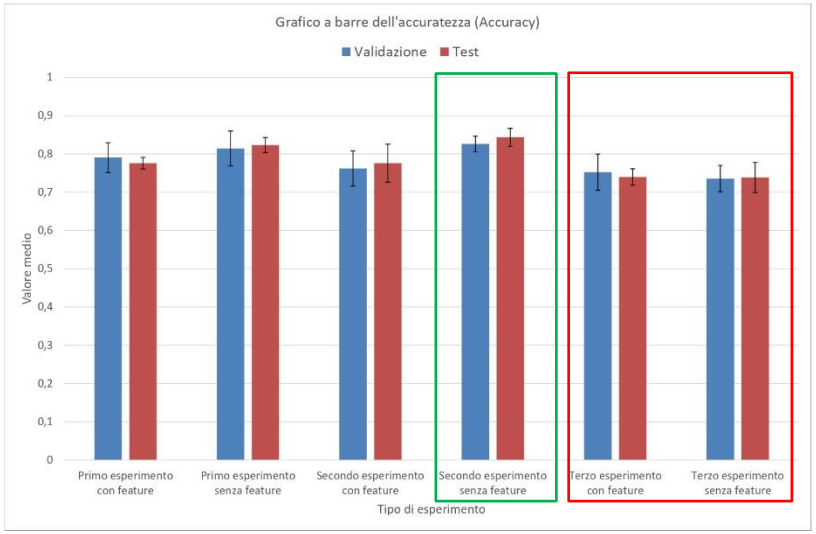
- Addestramento supervisionato
- Configurazione degli esperimenti: con e senza feature cliniche (race, ethnicity, age_at_diagnosis)
- K-fold cross-validation (stratificata), con k = 5
- Monitoraggio delle prestazioni: processo di apprendimento e generalizzazione finale sul test set
  - Loss, F1 weighted, Accuratezza, AUPRC, AUC

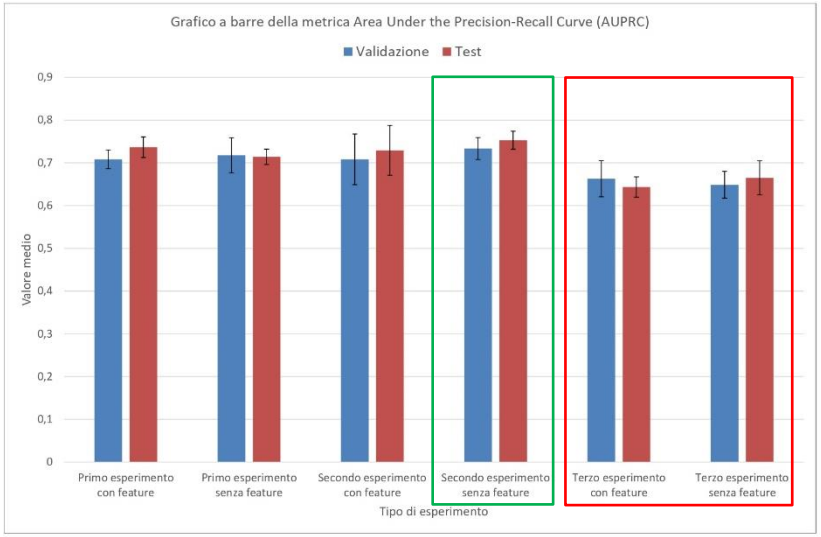
Risultati - F1 weighted

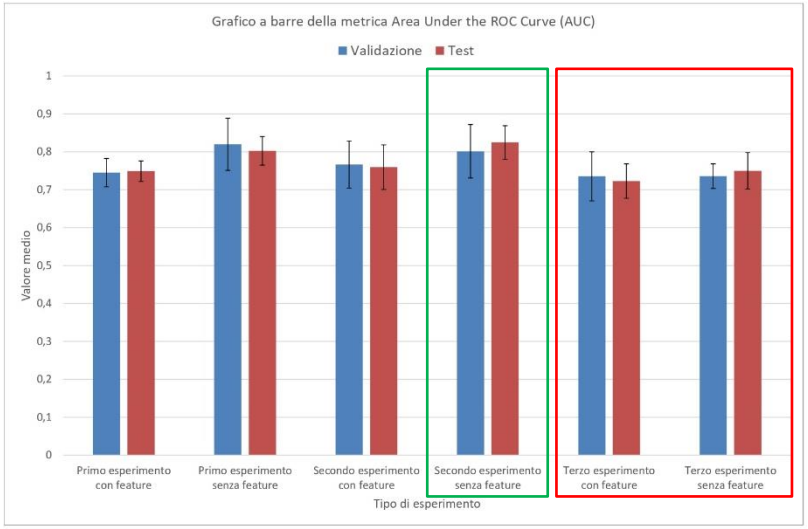
- La sparsificazione delle matrici unimodali e della matrice integrata migliora le performance
- La selezione delle feature tramite varianza ottiene i risultati peggiori
- Le feature cliniche peggiorano le performance



Risultati - Accuratezza, AUPRC, AUC







Confronto delle prestazioni sul dataset BRCA

- MOGDx, MOGONET e MoGCN sono metodi di integrazione multimodali basati su GCN

Metodo	N. Pazienti	Accuracy	F1-score
MOGDx	698	0.917 ± 0.028	0.901 ± 0.033
MOGONET	698	0.825 ± 0.006	0.816 ± 0.006
MoGCN	698	0.840 ± 0.024	0.851 ± 0.016
GraphSAGE (Esp. 2, no feat)	**696**	**0.843 ± 0,021**	**0.833 ± 0.023**

Tabella modificata da Ryan et al, Bioinformatics, 2023.

Conclusioni

- GraphSAGE efficace per modellare le relazioni tra pazienti
  - Valori di accuratezza, F1-score e AUC sempre superiori a 0,7
  - Risultati comparabili con i metodi allo stato dell'arte
- Feature cliniche non informative

- Sviluppi futuri
  - Ottimizzazione degli iperparametri del modello
  - Integrazione di ulteriori feature cliniche
  - Esplorare diverse funzioni di aggregazione (max-pooling o LSTM)
  - Estendere lo studio a diverse patologie (tumorali e non tumorali)

Grazie per l’attenzione