

UN MODELLO BASATO SU GRAPH CONVOLUTIONAL NETWORK PER LA STRATIFICAZIONE DI PAZIENTI CON PATOLOGIE NON TRASMISSIBILI



Dott. Francesco Grasso



Università
di Catania

INTRODUZIONE E OBIETTIVO



Contesto


- Neoplasia mammaria: sfida sanitaria primaria femminile
- La variabilità dei casi clinici necessita di approcci innovativi per categorizzare i pazienti



Obiettivo

Sviluppare un sistema AI di tipo GCN (Graph Convolutional Networks) per la previsione di outcomes clinici su dati METABRIC

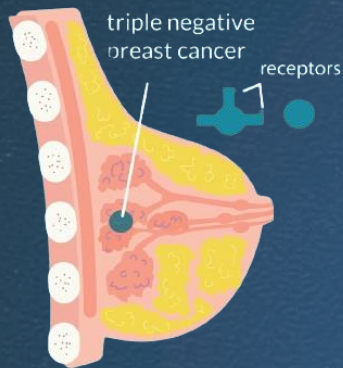




GRAFI E GRAPH CONVOLUTIONAL NETWORKS (GCN)

- Grafo non orientato
- GCN
- dati con forma irregolare





COS' È IL TNBC

Tipo di cancro al seno caratterizzato dall'assenza dei tre recettori comuni:

- Recettore degli estrogeni (ER)
- Recettore del progesterone (PR)
- Recettore 2 del fattore di crescita epidermico umano (HER2)

Rappresenta circa il 10-20% di tutti i casi di cancro al seno

Generalmente ha una prognosi peggiore e meno opzioni terapeutiche mirate.



DATASET METABRIC

Il dataset METABRIC è una risorsa fondamentale nella ricerca sul cancro al seno, incluso il Triple Negative Breast Cancer (TNBC).

Creato dal consorzio internazionale Molecular Taxonomy of Breast Cancer, fornisce un'analisi molecolare dettagliata dei tumori al seno per migliorare la classificazione, diagnosi e trattamento.

Ecco le componenti del Dataset:

- **Dati Clinici:** Informazioni dettagliate su oltre 2.000 pazienti con cancro al seno, tra cui dati demografici, caratteristiche del tumore, stadio del cancro, trattamenti ricevuti e outcome clinici.
- **Dati Genomici:** Profilazione genetica completa utilizzando tecniche come la sequenza del DNA, RNA-seq. Include dati sulle mutazioni, espressione genica, e variazioni nel numero di copie.

Dati Specifici per il TNBC:

- **Esiti Clinici:** Dati sui tassi di sopravvivenza, risposta ai trattamenti e ricorrenza del cancro per pazienti TNBC.
- **Profilazione Genomica:** Utilizzata per identificare nuovi sottotipi molecolari di TNBC e per comprendere meglio la biologia di questo tumore aggressivo.

DATASETS UTILIZZATI

PATIENT_ID	94	2334	392636	186	79026	113146		7402	677	OS	OS.time	level
MB-0062	0	0	0	0	0	0		0	0	0	12.83	OS>=2
MB-0079	0	0	1	0	0	0		0	0	1	2.38	OS>=2
MB-0100	0	0	0	0	0	0		0	0	1	0.67	OS<2
MB-0115	0	0	0	0	0	0		0	0	1	5.56	OS>=2
MB-0127	0	0	0	0	0	0	...	0	0	0	11.01	OS>=2
MB-0149	0	0	0	0	0	0		0	0	1	4.31	OS>=2
MB-0157	0	0	0	0	0	0		0	0	0	9.56	OS>=2
MB-0164	0	0	0	0	0	0		0	0	0	0.9	OS<2
MB-0174	0	0	0	0	0	0		0	0	0	6.56	OS>=2
MB-0188	0	1	0	0	0	0		0	0	1	2.61	OS>=2
⋮												
MB-5634	0	0	0	0	0	0		0	0	1	5.24	OS>=2
MB-5651	0	0	0	0	0	0		0	0	1	1.75	OS<2
MB-5655	0	0	0	0	0	0		0	0	0	15.98	OS>=2
MB-6143	0	0	0	0	0	2	...	0	0	0	21.3	OS>=2
MB-6223	0	0	0	0	0	0		0	0	0	18.49	OS>=2
MB-6237	0	0	0	0	0	0		0	1	1	8.77	OS>=2
MB-6251	0	0	0	0	0	0		0	0	1	1.22	OS<2



DATASETS UTILIZZATI

PATIENT_ID X1	X29974	X2	X127550	X53947	X51146		X23140	X26009	OS	OS.time	level
MB-0062	5,412713	5,57146	8,675327	5,372409	6,12862	5,624841	5,912904	9,40027	0	12,83	OS>=2
MB-0079	5,163773	5,412081	10,67962	5,233153	5,723339	5,218649	6,678373	8,959534	1	2,38	OS>=2
MB-0100	5,204053	5,414651	9,744305	5,43954	6,636901	5,788499	7,702676	8,475007	1	0,67	OS<2
MB-0115	5,612779	5,312095	10,88449	5,214144	6,206407	5,401816	7,377194	8,118007	1	5,56	OS>=2
MB-0127	5,458105	5,392807	9,843684	5,077362	6,542352	5,431338	7,334331	8,121314	0	11,01	OS>=2
MB-0149	5,367511	5,509567	9,681886	5,176954	5,825413	5,291021	8,396715	8,033341	1	4,31	OS>=2
MB-0157	5,58462	5,051621	10,47546	4,930657	7,189692	5,709059	7,474057	7,913059	0	9,56	OS>=2
MB-0164	5,399935	5,435363	9,596402	5,32333	6,265992	5,368745	7,170153	8,045316	0	0,9	OS<2
MB-0174	5,578841	5,714959	10,45274	5,325357	7,783416	5,716028	8,015598	7,693857	0	6,56	OS>=2
MB-0179	5,372652	5,371565	10,62555	5,254758	7,071885	5,463147	7,558208	7,873357	1	1,49	OS<2
MB-0188	5,551104	5,462855	10,5201	5,403241	6,109512	5,678003	7,380718	8,096481	1	2,61	OS>=2
⋮											
MB-5577	5,560725	5,328761	10,56704	5,329585	6,069552	5,670466	8,610783	8,132073	0	15,19	OS>=2
MB-5616	5,539039	5,450282	9,818055	5,056321	6,346233	5,556551	7,367912	8,267504	0	15,55	OS>=2
MB-5633	5,509044	5,348039	10,29152	5,093135	6,284927	5,563549	8,046165	8,306895	0	14,89	OS>=2
MB-5634	5,507712	5,511096	10,63361	5,348715	5,990919	5,56883	7,947798	8,146131	1	4,62	OS>=2
MB-5651	5,376763	5,263582	9,100757	5,298644	6,44042	5,307602	7,737826	8,752925	0	15,05	OS>=2
MB-5655	5,55551	5,384084	10,46553	5,194813	6,493361	5,678821	8,20563	8,40266	1	2,92	OS>=2
MB-6143	5,424203	5,224852	11,60822	5,08454	6,744245	5,40881	7,731763	8,388021	1	5,24	OS>=2
MB-6223	5,366654	5,344031	9,046238	5,258917	6,057556	5,250579	7,662384	8,729026	1	1,75	OS<2
MB-6237	5,633888	5,142537	7,884312	4,935283	7,281366	5,720547	7,436167	8,885672	0	15,98	OS>=2
MB-6251	5,424336	5,234556	7,78792	5,174407	5,946628	5,742796	7,651138	7,565239	0	21,3	OS>=2

IL NOSTRO APPROCCIO

LE TRE FASI:

Pre-processing dei dati

generazione di
distribuzioni
normalizzate per ogni
paziente e ogni gene

Creazione della rete

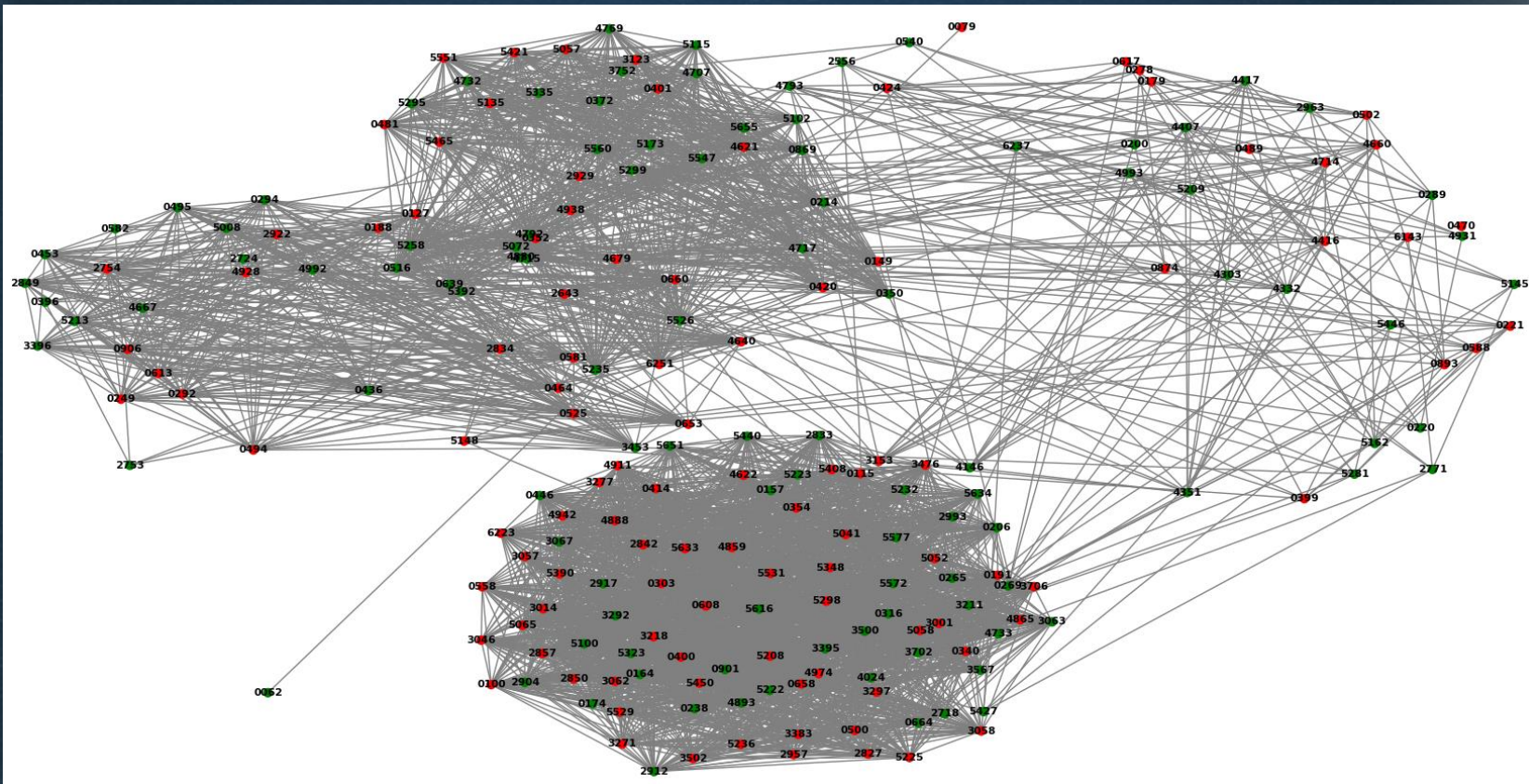
l'algoritmo K-Means

Classificazione tramite GCN

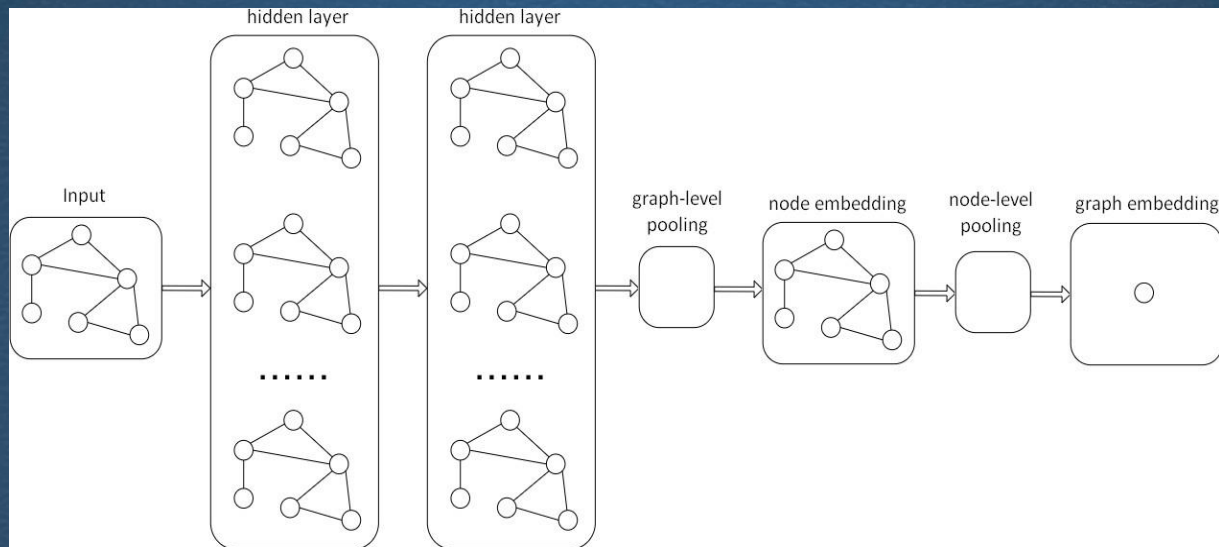
Classi di convoluzione
Struttura del modello
GCN (input, strati
nascosti, output)



LA RETE DI PAZIENTI



CONVOLUZIONE E STRUTTURA GCN



$$a \mathbf{h}_v^k = \sigma \left(\mathbf{w}_k \cdot \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{b}_k \mathbf{h}_v^{k-1} \right)$$

ADDESTRAMENTO E TEST DEL MODELLO



Train e validation

- Forward propagation
- Calcolo della perdita (Loss)
- Backpropagation e
aggiustamento dei parametri



Test

- Forward propagation
- Calcolo della perdita (Loss)
- Calcolo delle metriche di
valutazione



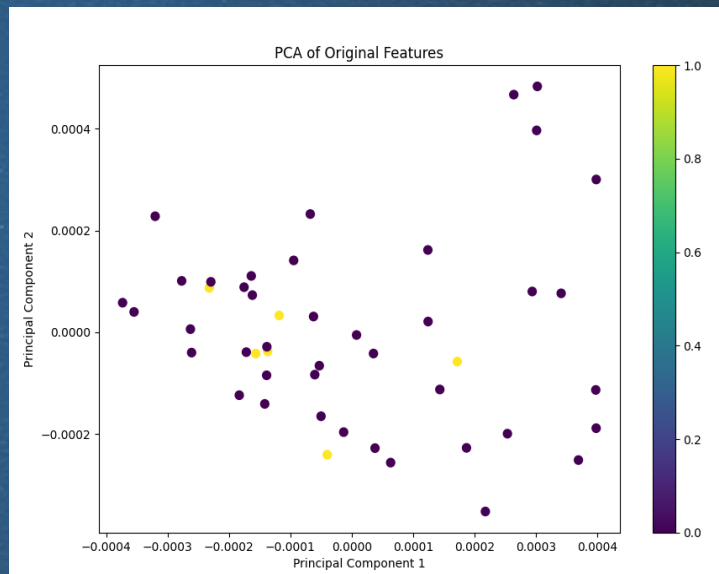
BILANCIAMENTO DATASET



SMOTE

Key performance indicator	Number
Campioni della classe minoritaria	30
Campioni dopo l'applicazione dello SMOTE	183
Campioni totali del dataset	366

Grafico dei pazienti



SPERIMENTAZIONE SUGLI OTTIMIZZATORI

Ottimizzatore	Max train accuracy	Min train loss	Test accuracy	Test loss
<u>Adam</u>	90%	0,26	93,4%	0,20
<u>Adam W</u>	88,4%	0,25	90,4%	0,22
<u>RMSProp</u>	87,3%	0,25	88,4%	0,23
<u>Adagrad</u>	91,2%	0,24	94,4%	0,19



SPERIMENTAZIONE SULLE FUNZIONI DI ATTIVAZIONE

F. di attivazione	Max train accuracy	Min train loss	Test accuracy	Test loss
<u>Log_softmax</u>	90%	0,26	92,4%	0,21
<u>Softmax</u>	85,4%	0,3	83,4%	0,31
<u>Sigmoid</u>	85,3%	0,3	88,8%	0,31
<u>Tanh</u>	83,2%	0,33	84,4%	0,35
<u>Softplus</u>	88,3	0,28	85,5	0,25



NORMALIZZAZIONE DELLA MATRICE DI ADIACENZA

la normalizzazione della matrice di adiacenza nella funzione forward della GCN ha rappresentato un punto di svolta, migliorando significativamente le prestazioni del modello.

Con questa modifica abbiamo ottenuto come miglioramenti:

- Accuracy di training dal 90,03% al 94,58%
- Loss di training da 0,266 a 0,149

RISULTATI FINALI E CONFRONTO CON ALTRI MODELLI

GCN

- Loss= 0.097
- Accuracy= 97%
- F1= 96%
- Precision= 96%
- Recall= 96%

GraphSage

- Loss= 0.03
- Accuracy= 99%

GAT

- Loss= 0.69
- Accuracy= 50%

Risultati fase di Test dei tre modelli:

- Loss= 0.11
- Accuracy= 94%
- F1= 94%
- Precision= 94%
- Recall= 94%

**GRAZIE PER
L'ASCOLTO!**

