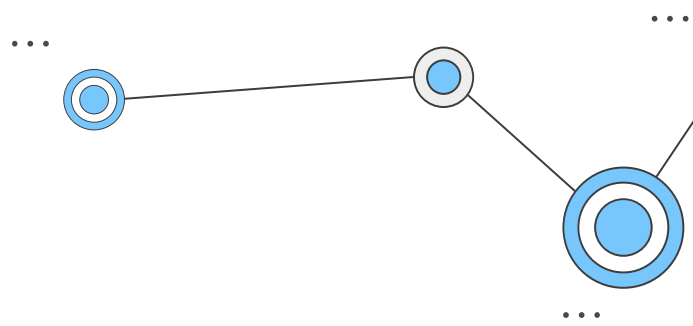




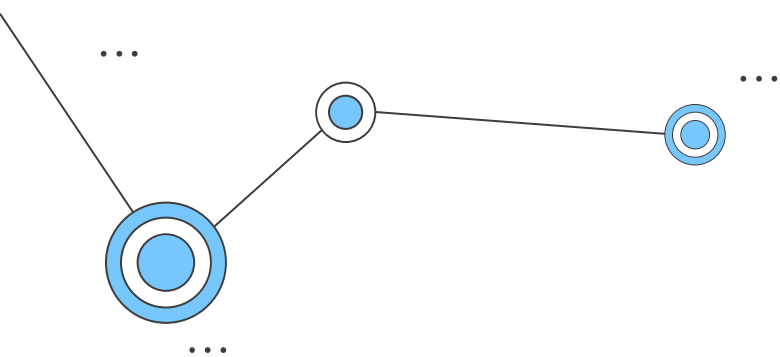
Candidato:
Lemuel Puglisi

Relatore:
Prof. Alaimo
Salvatore

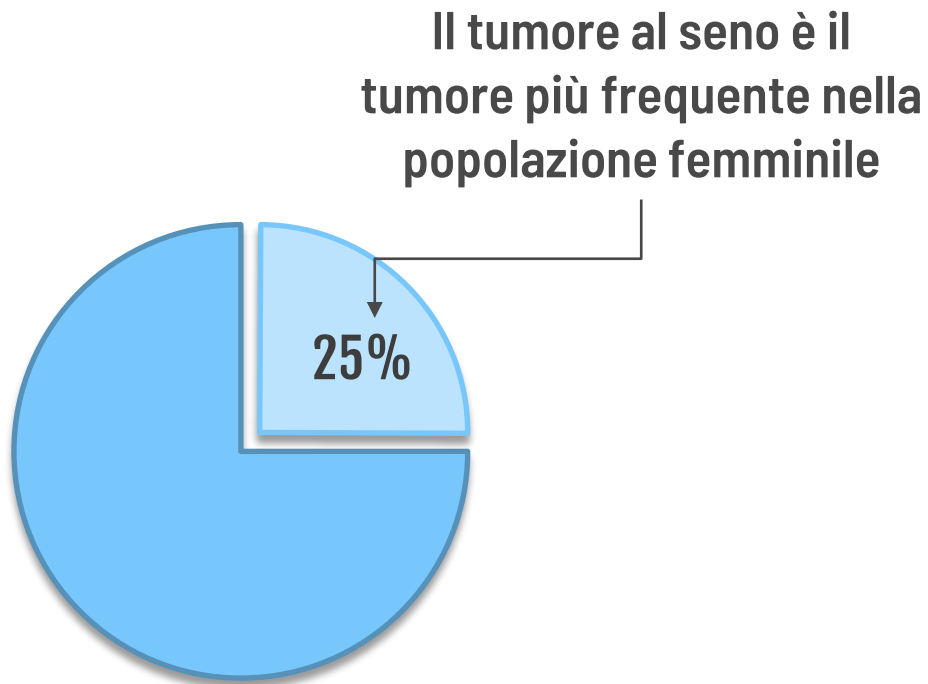
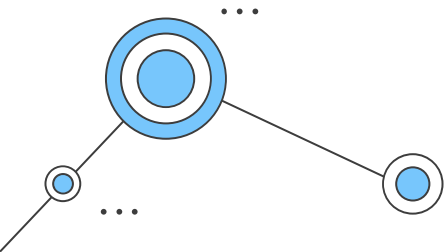
Correlatori:
Prof. Ferro Alfredo
Dott. Micale Giovanni



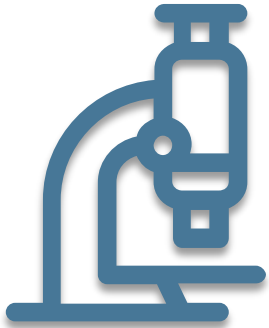
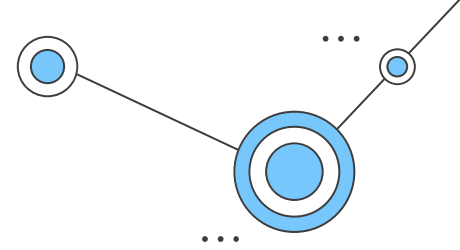
Autoencoder per la riduzione della dimensionalità di dataset molecolari e conseguente predizione di dati clinici



Contesto: Tumore al seno



Sottotipi molecolari



Lo studio a livello molecolare ha portato alla definizione di sottotipi molecolari di carcinoma mammario, la cui distinzione favorisce l'adozione di **terapie mirate**.

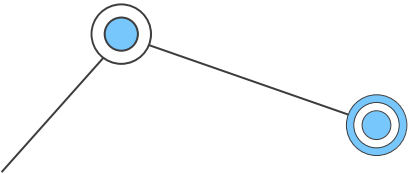
Luminal A

Luminal B

Her2+

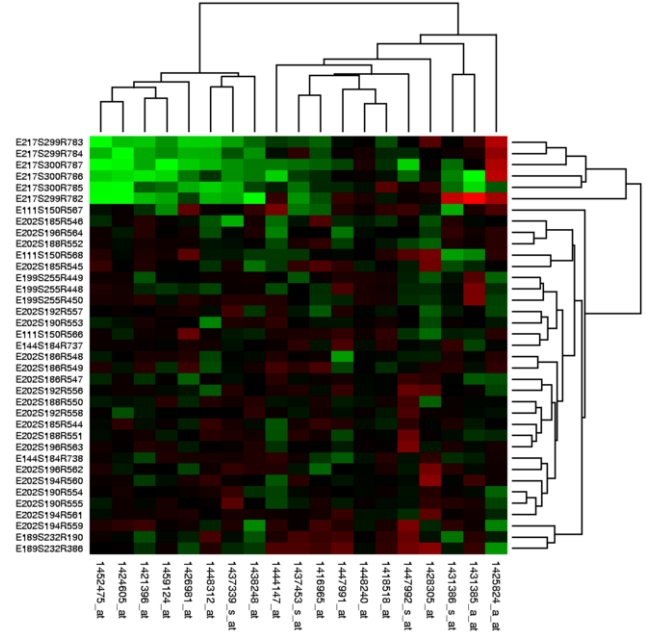
Basal

E molti altri...



Profilo dell'espressione genica

Un profilo
dell'espressione
genica rappresenta
l'attività di migliaia
di geni in un dato
istante.



Heatmap by Miguel Andrade

Genes expression profiling

Per ottenere profili dell'espressione genica dal tessuto ammalato, vengono utilizzate varie tecniche. I dataset forniti per lo studio adottano rispettivamente le tecniche RNA-Seq e Microarray.

RNA-Seq

Tecnica NGS.
Sequenziamento dell'RNA
e allineamento su genoma

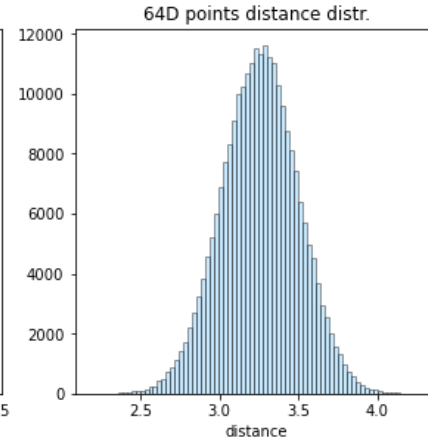
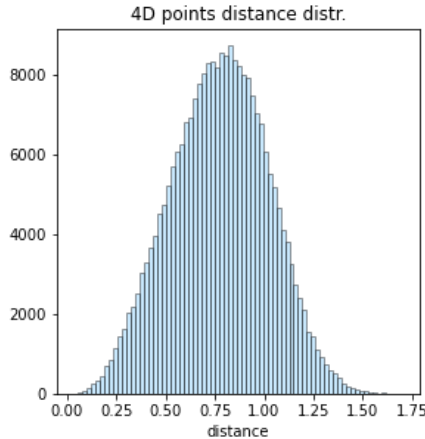
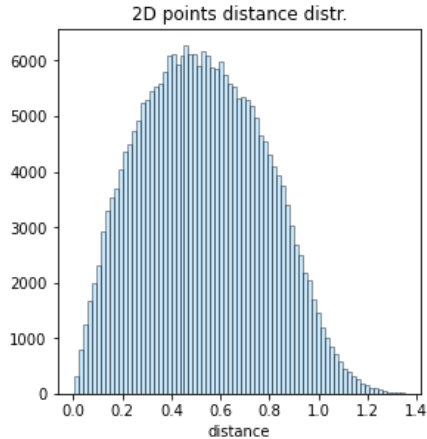


Microarray

Sonde microscopiche
analizzano simultaneamente
la presenza di geni

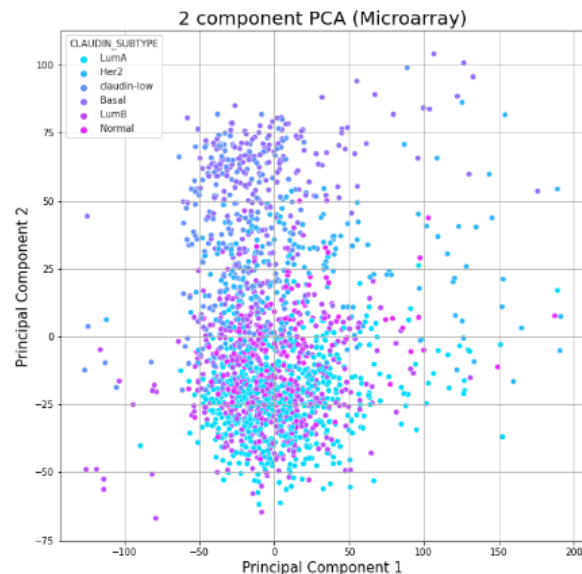
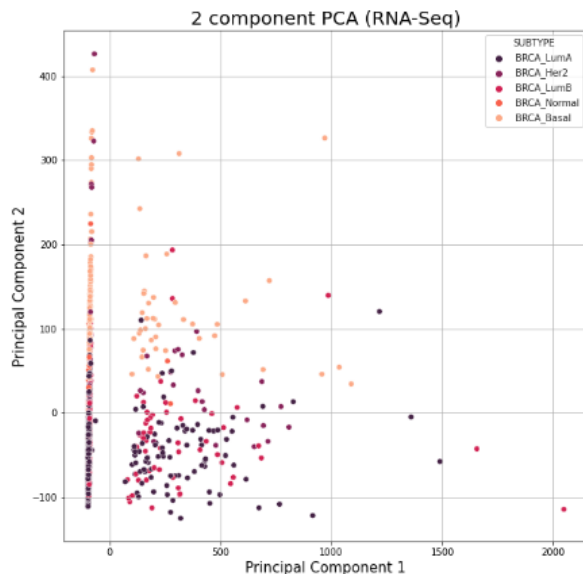
Curse of dimensionality

Ogni profilo dell'espressione genica è un dato ad alta dimensionalità: si fa riferimento a circa 20.000 geni del genoma umano. L'alta dimensionalità dei dati provoca il fenomeno della **curse of dimensionality**: il volume dello spazio aumenta esponenzialmente con le dimensioni, per cui ad alte dimensioni i punti risultano pressoché **equidistanti**.

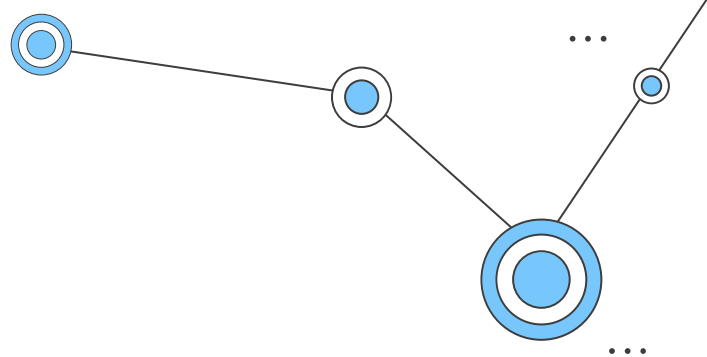


Principal Component Analysis

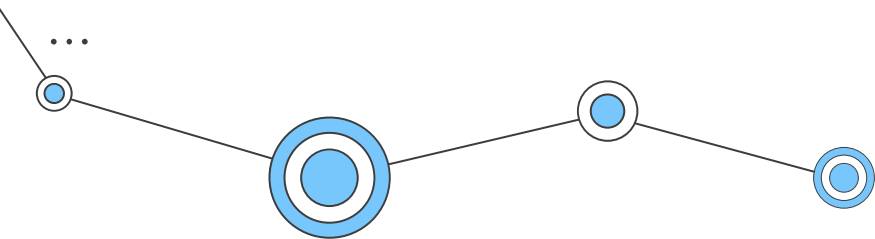
Metodi classici di riduzione della dimensionalità, come la PCA, performano male quando vi sono relazioni non lineari tra le feature.



Perché utilizzare delle **reti neurali** per la riduzione della dimensionalità?

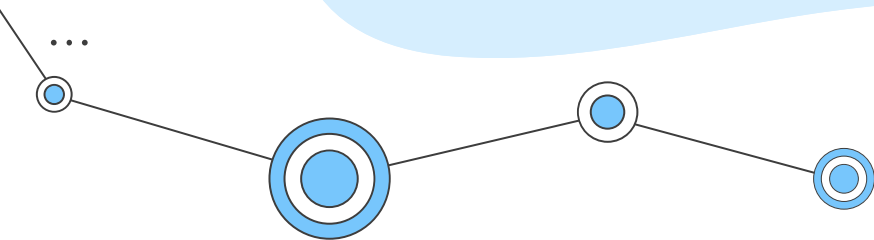
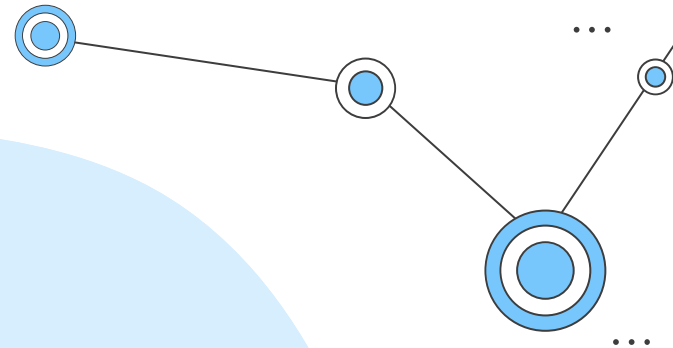


1. Le reti neurali sono **approssimatori universali**.
2. Le performance aumentano proporzionalmente alla **quantità di dati**

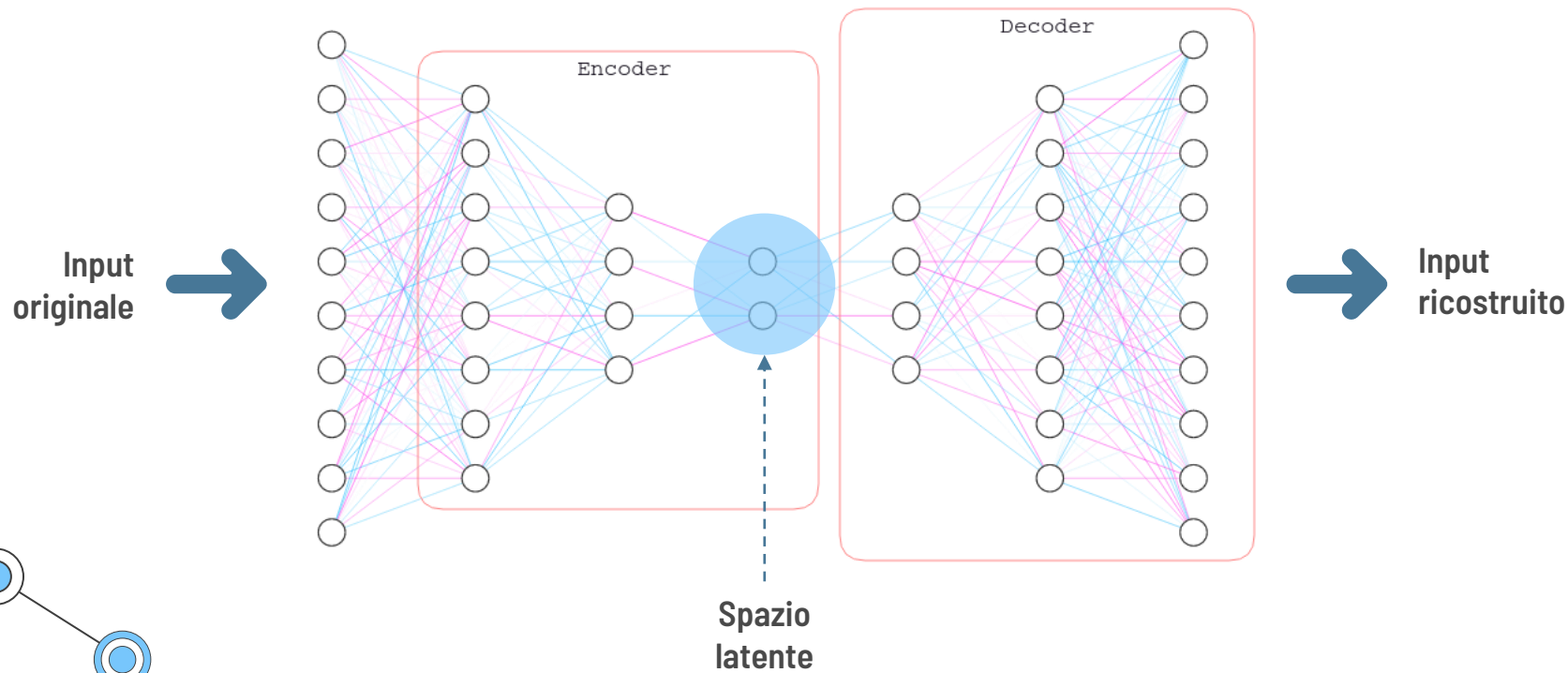




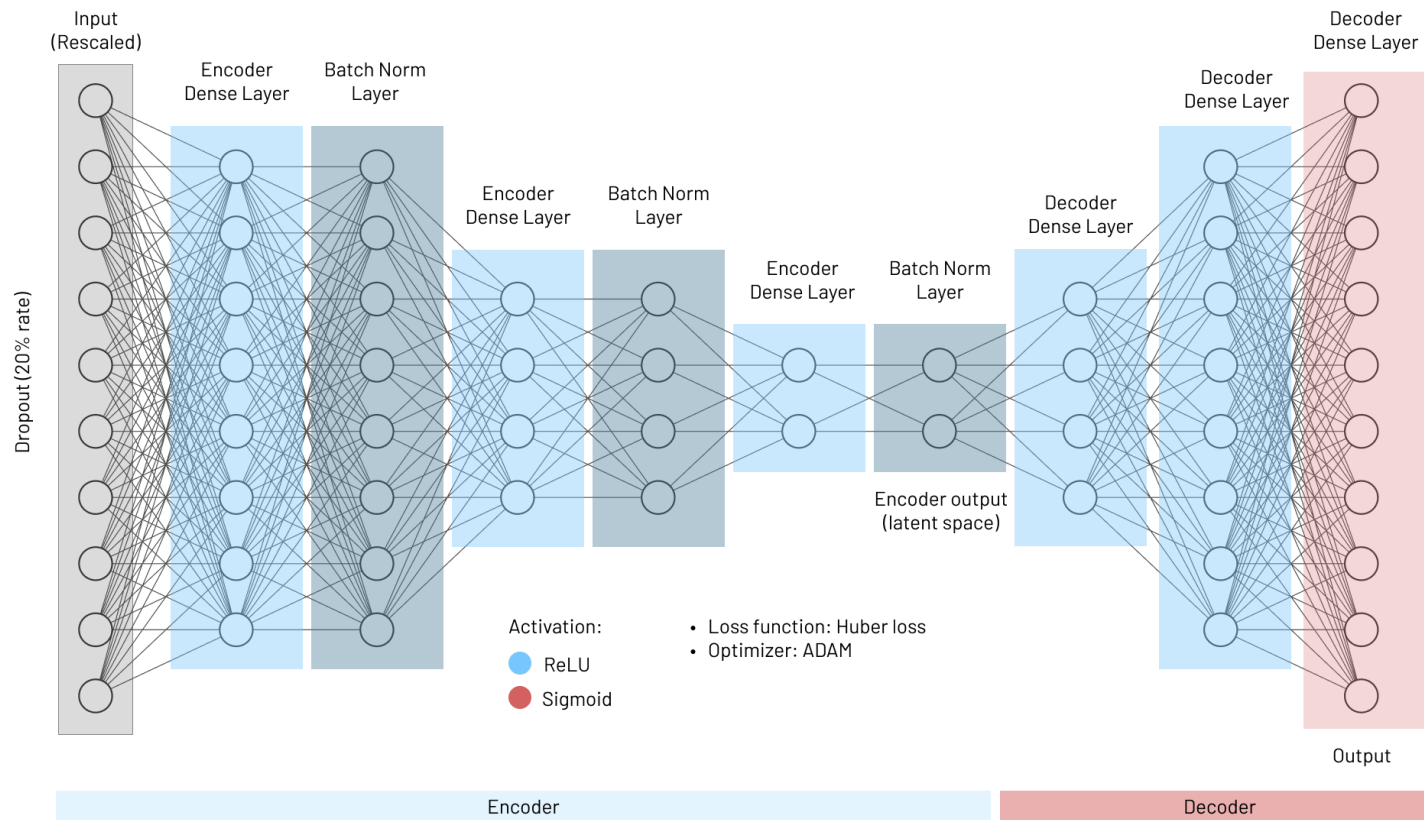
Autoencoders

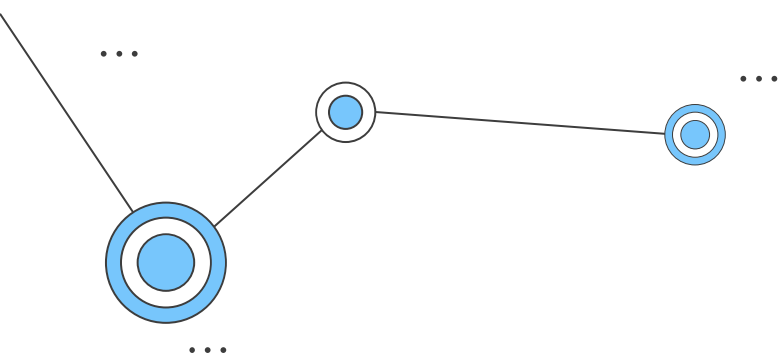


Struttura di un deep autoencoder



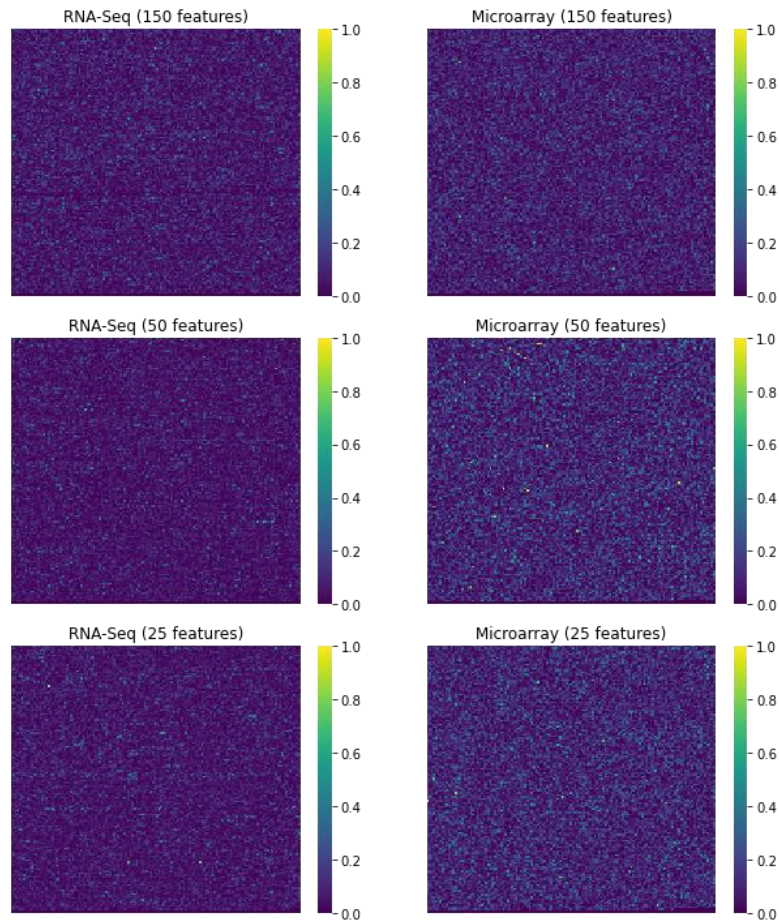
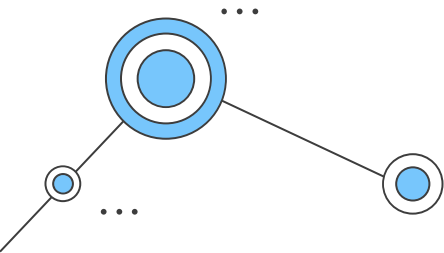
Scelte implementative






Risultati di compressione

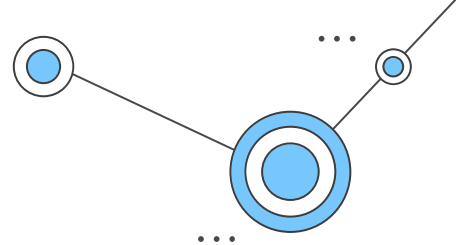
La huber loss media si aggira a 0.008
nei dati RNA-Seq e 0.013 per i dati
Microarray.





Classificazione del sottotipo attraverso i dati compressi

Modelli di classificazione



XGBoost

Modello ensemble, utilizza
il gradient boosting



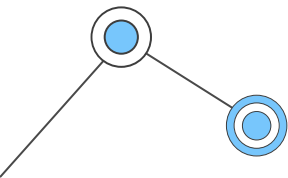
Random Forest

Modello ensemble,
utilizza il bootstrapping



SVM

Modello discriminativo,
sfrutta vettori di supporto



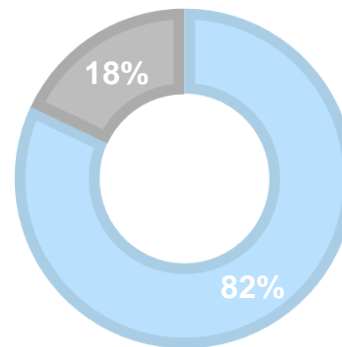
* Ci riferiremo al modello XGBoost quando parleremo di risultati, poiché le performance sono lievemente migliori.

Risultati su dati RNA-Seq compressi

Embedding space	Metrics	LumA	LumB	Basal	Her2
150	Sensitivity	0.94	0.72	0.97	0.44
	Specificity	0.89	0.91	0.99	0.99
	Precision	0.90	0.67	0.97	0.78
	NPV	0.93	0.93	0.99	0.95
50	Sensitivity	0.91	0.69	0.97	0.38
	Specificity	0.83	0.91	0.99	0.98
	Precision	0.86	0.68	0.97	0.67
	NPV	0.89	0.92	0.99	0.94
25	Sensitivity	0.90	0.54	0.91	0.38
	Specificity	0.83	0.89	0.99	0.96
	Precision	0.86	0.55	0.94	0.46
	NPV	0.88	0.88	0.98	0.94

ACCURATEZZA

■ Corrette ■ Errate



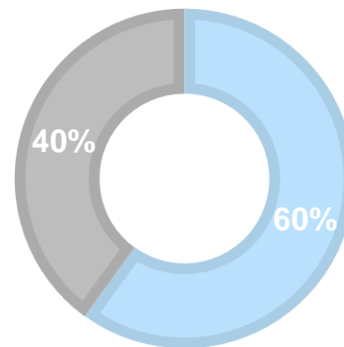
* Ci riferiremo al modello XGBoost quando parleremo di risultati, poiché le performance sono lievemente migliori.

Risultati su dati Microarray compressi

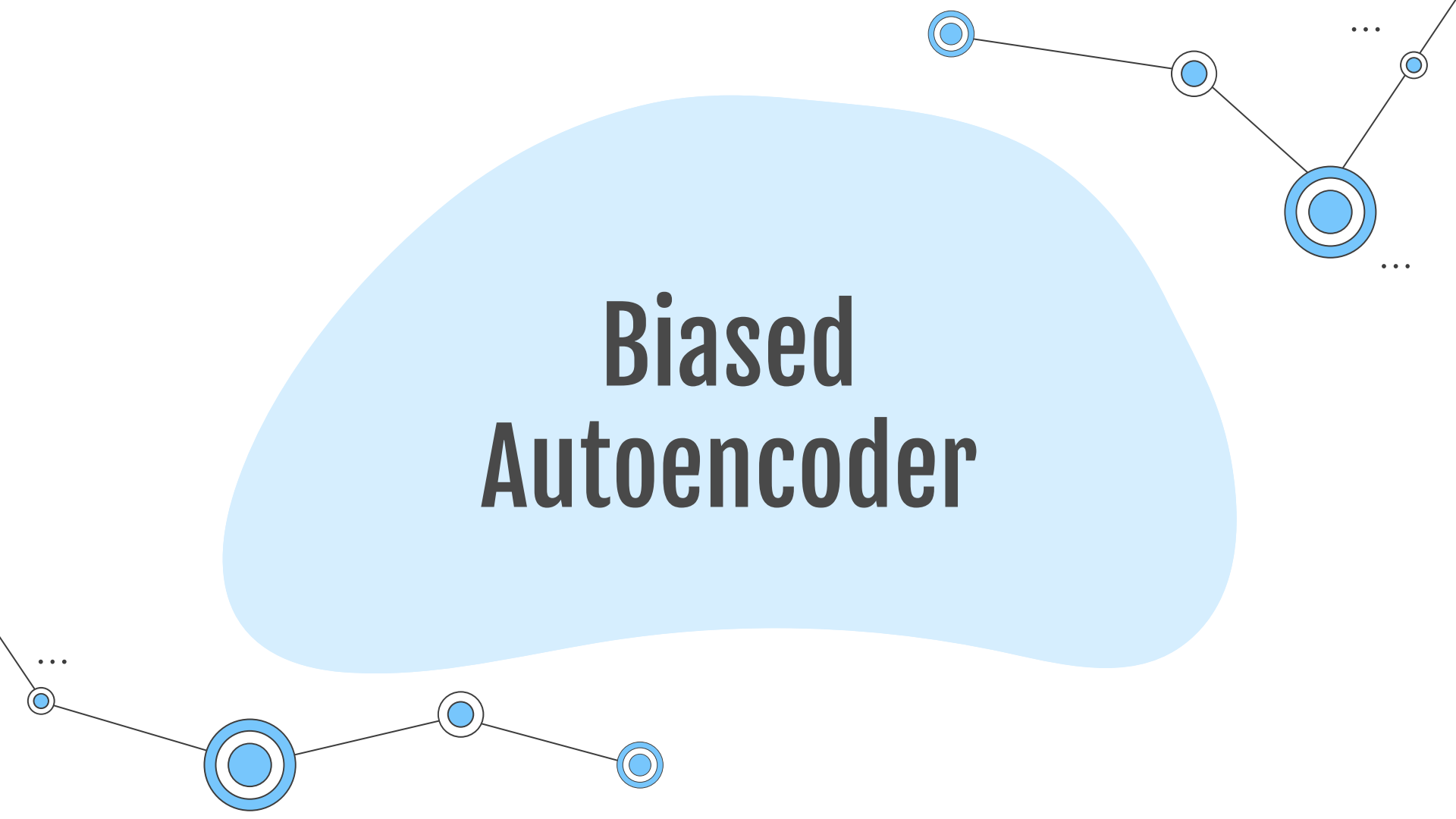
Embedding space	Metrics	LumA	LumB	Basal	Her2
150	Sensitivity	0.60	0.72	0.92	0.77
	Specificity	0.95	0.78	0.98	0.89
	Precision	0.90	0.58	0.86	0.53
	NPV	0.76	0.87	0.99	0.96
50	Sensitivity	0.67	0.68	0.95	0.68
	Specificity	0.93	0.82	0.96	0.90
	Precision	0.88	0.62	0.79	0.52
	NPV	0.78	0.86	0.99	0.94
25	Sensitivity	0.66	0.67	0.90	0.68
	Specificity	0.92	0.81	0.97	0.88
	Precision	0.87	0.60	0.84	0.49
	NPV	0.78	0.86	0.99	0.94

ACCURATEZZA

■ Corrette ■ Errate

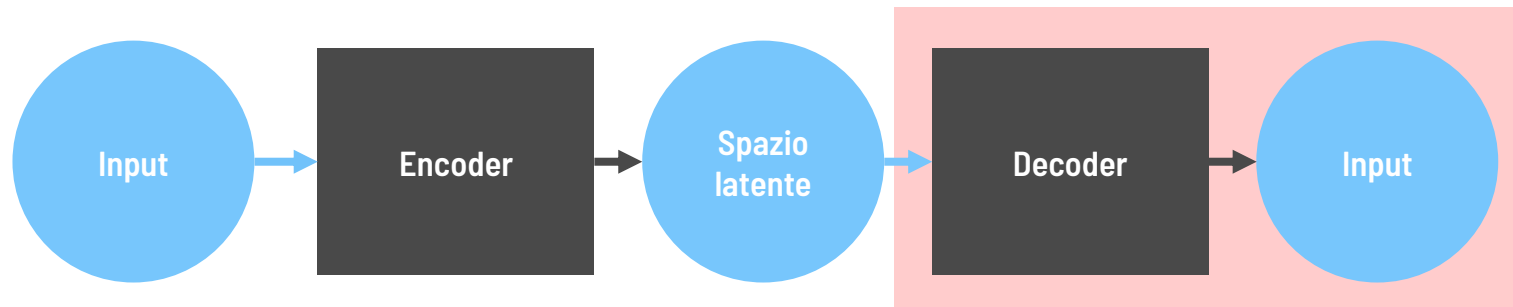


Biased Autoencoder

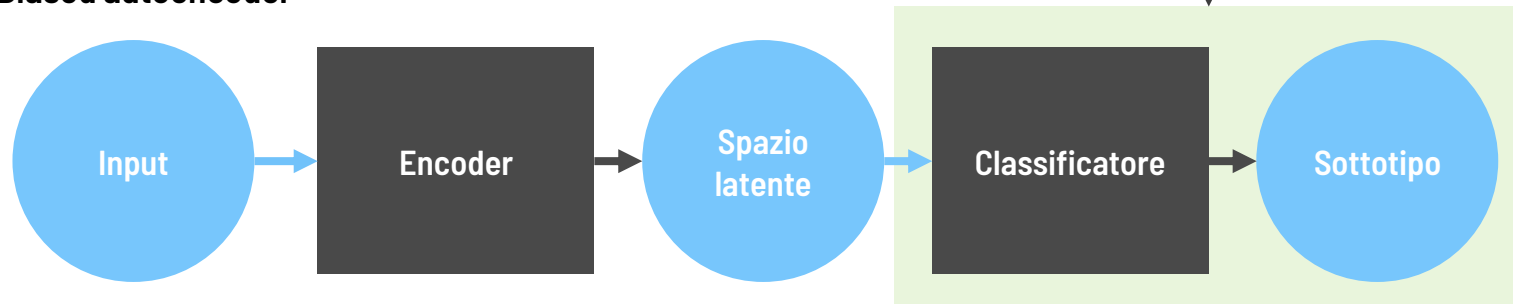


Idea di base

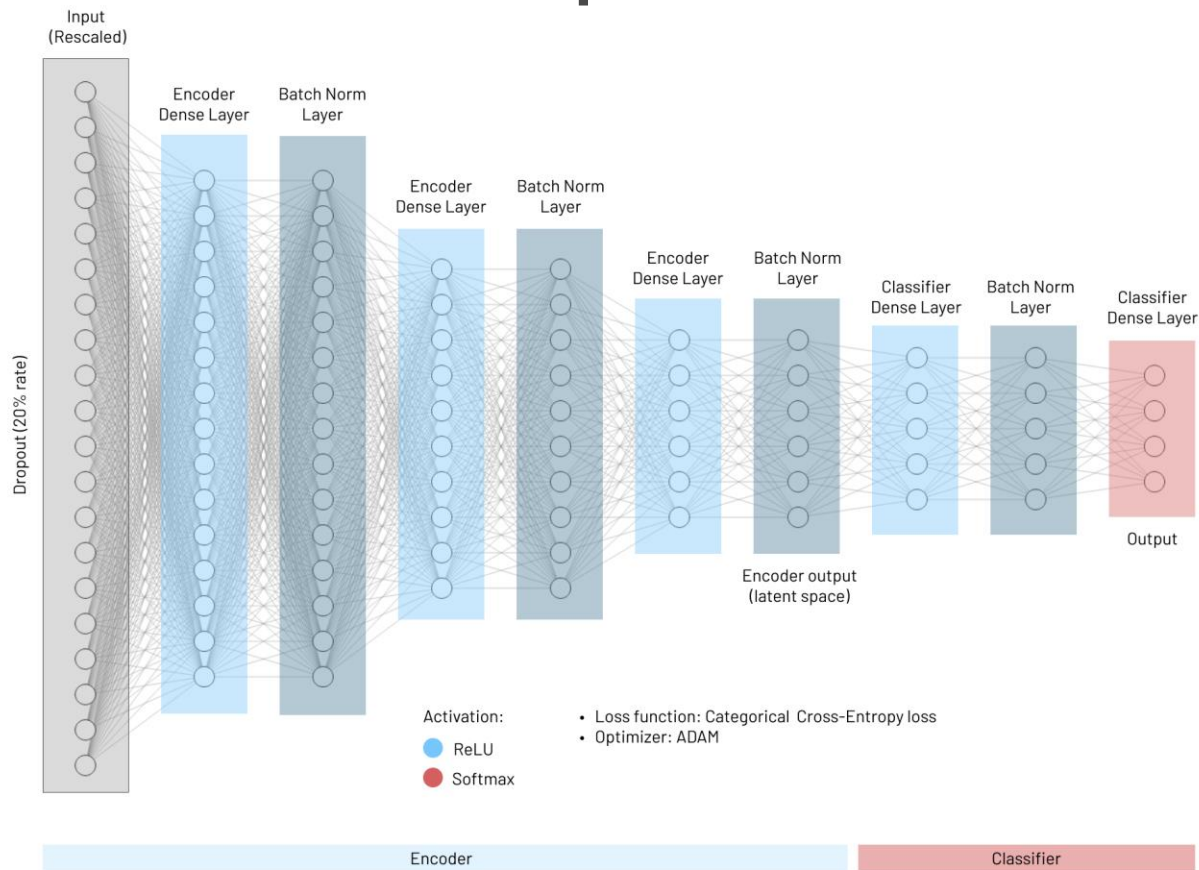
Autoencoder classico

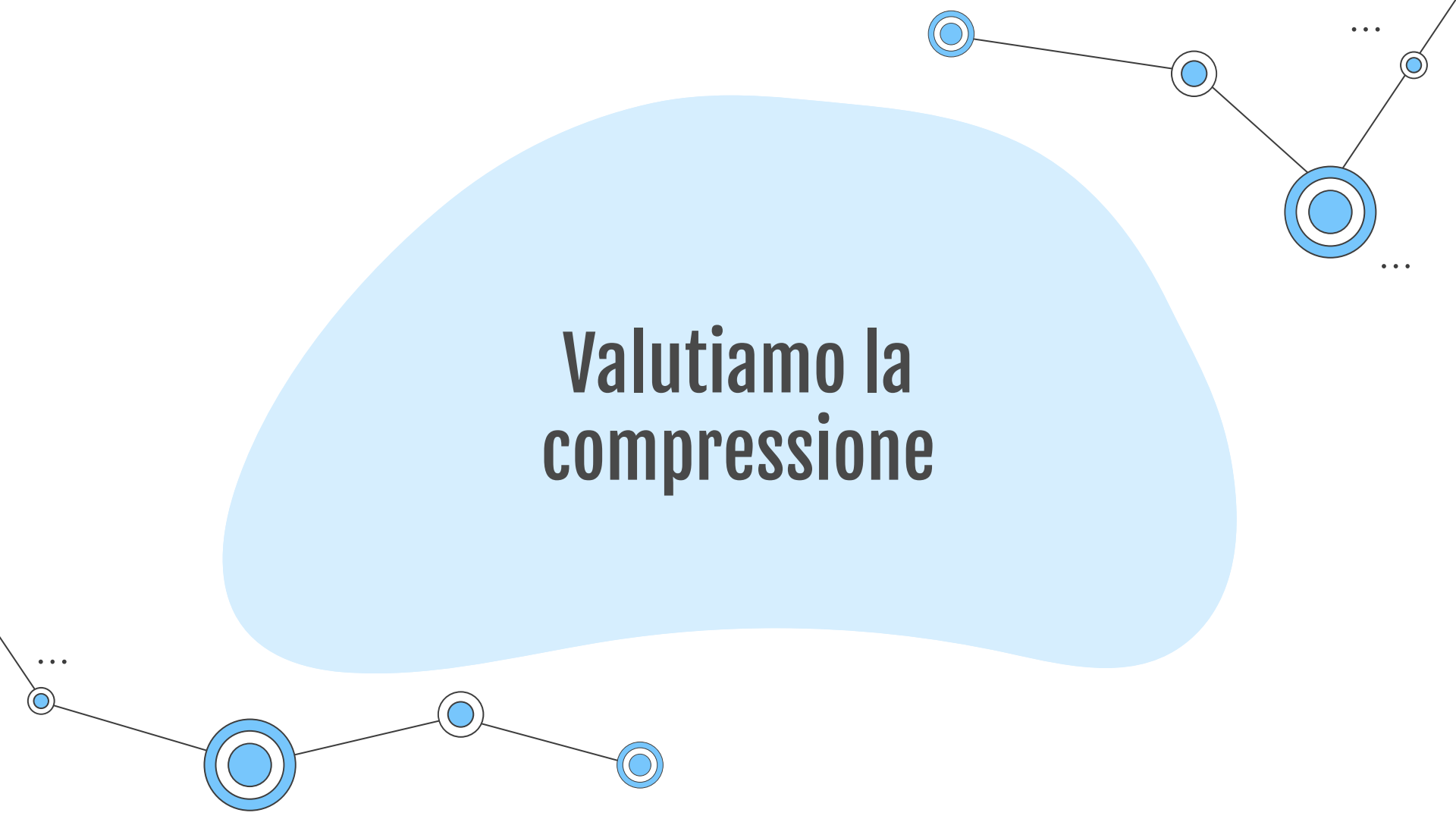


Biased autoencoder



Scelte implementative





**Valutiamo la
compressione**

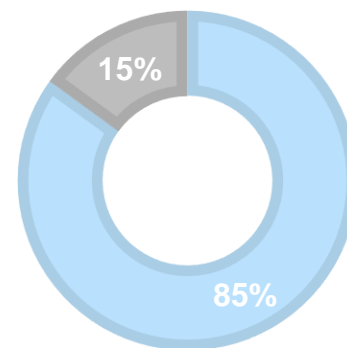
* Ci riferiremo al modello XGBoost quando parleremo di risultati, poiché le performance sono lievemente migliori.

Risultati su dati RNA-Seq compressi (A.E. biased)

Embedding space	Metrics	LumA	LumB	Basal	Her2
150	Sensitivity	0.96	0.90	0.97	0.94
	Specificity	0.98	0.97	1.00	0.98
	Precision	0.98	0.90	1.00	0.79
	NPV	0.96	0.97	0.99	0.99
50	Sensitivity	0.98	0.79	1.00	0.94
	Specificity	0.94	0.99	0.99	0.98
	Precision	0.95	0.94	0.97	0.83
	NPV	0.98	0.95	1.00	0.99
25	Sensitivity	0.98	0.77	1.00	0.88
	Specificity	0.93	0.99	0.99	0.97
	Precision	0.94	0.97	0.97	0.74
	NPV	0.98	0.94	1.00	0.99

ACCURATEZZA

■ Corrette ■ Errate



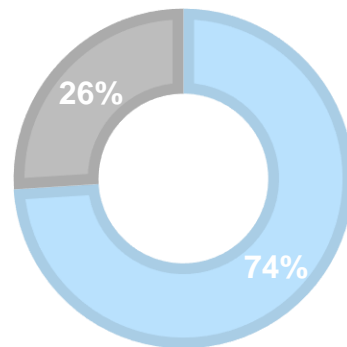
* Ci riferiremo al modello XGBoost quando parleremo di risultati, poiché le performance sono lievemente migliori.

Risultati su dati Microarray compressi (A.E. biased)

Embedding space	Metrics	LumA	LumB	Basal	Her2
150	Sensitivity	0.83	0.89	0.98	0.91
	Specificity	0.97	0.91	0.98	0.97
	Precision	0.96	0.80	0.89	0.83
	NPV	0.88	0.95	1.00	0.98
50	Sensitivity	0.85	0.87	0.98	0.89
	Specificity	0.97	0.91	0.97	0.97
	Precision	0.95	0.81	0.85	0.85
	NPV	0.89	0.94	1.00	0.98
25	Sensitivity	0.87	0.90	1.00	0.89
	Specificity	0.97	0.91	0.98	0.97
	Precision	0.96	0.81	0.87	0.83
	NPV	0.87	0.96	1.00	0.98

ACCURATEZZA

■ Corrette ■ Errate



Differenze tra i due modelli

Blind autoencoder

Produce una
rappresentazione
generica del dato,
utilizzabile per
qualsiasi tipo di analisi.

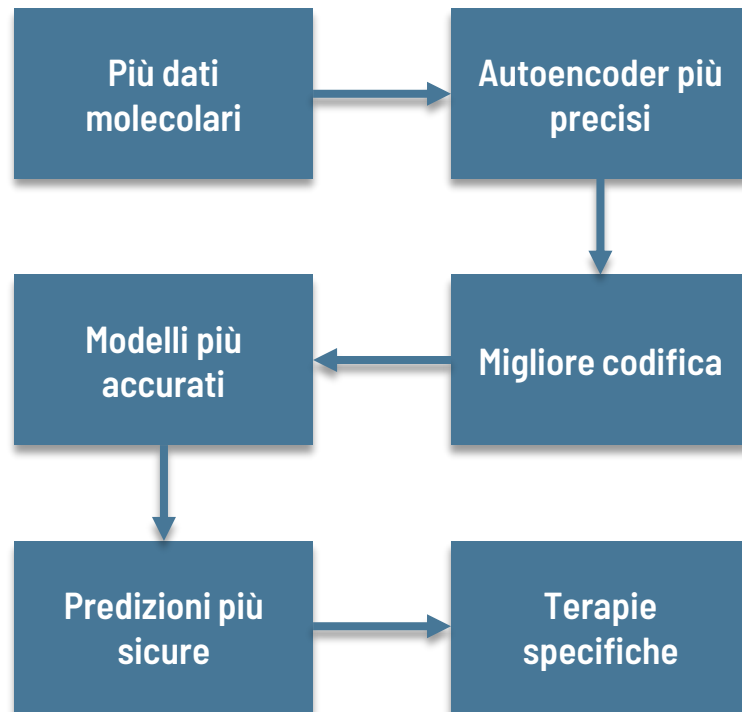
Biased autoencoder

La rappresentazione
prodotta è basata su
una determinata
proprietà dei dati, su
cui è effettuato il
training del modello.



Conclusione

La progressiva quantità di dati sul carcinoma mammario aiuta la produzione di autoencoder precisi, quindi compressioni significative. I modelli di predizione lavorano su dati a bassa dimensionalità e producono risultati più accurati. I modelli possono essere sfruttati come strumenti di supporto alle decisioni, per lo sviluppo di trattamenti mirati.



**Grazie
dell'attenzione**

