Using a semi-supervised method to identify breast cancer patients with similar characteristics

Vesna Cuplov

LITO, Institut Curie, Inserm, U1288, Orsay, France

October 6th 2021

Reminder about my post-doc project

- **PAN**omic **A**tlas for non-small **CE**ll lung cancer manag**E**ment

- Develop methods & tools to **identify** a small **group of patients** with non small cell lung cancer and **similar clinical and radiomic characteristics**

- This small group of patients would be extracted from a reference database (under construction: 58 patients so far)

Lung cancer cohort

- The medical history of these **“twin-patients”** will allow doctors to suggest the therapeutic strategy to be adopted for a new patient

Patients and image acquisition

- While waiting to increase the RALUCA-lung database, we test our methods on the **RALUCA-breast database** composed of **289 patients**
- Radiomic features were extracted from the breast **primary tumor** (using a 40%SUVmax threshold) and on a **ring around the tumor**
- Radiomic features were extracted from a baseline PET scan using the **LIFEx** software
- Several **clinical parameters** were collected: Age, T/N/M stage, BMI, Menopause status, Hormon receptors: progesterone receptor (PR), estrogen receptor (ER), human epidermal growth factor receptor 2 (HER2) and the nuclear protein Ki-67 (antigen)

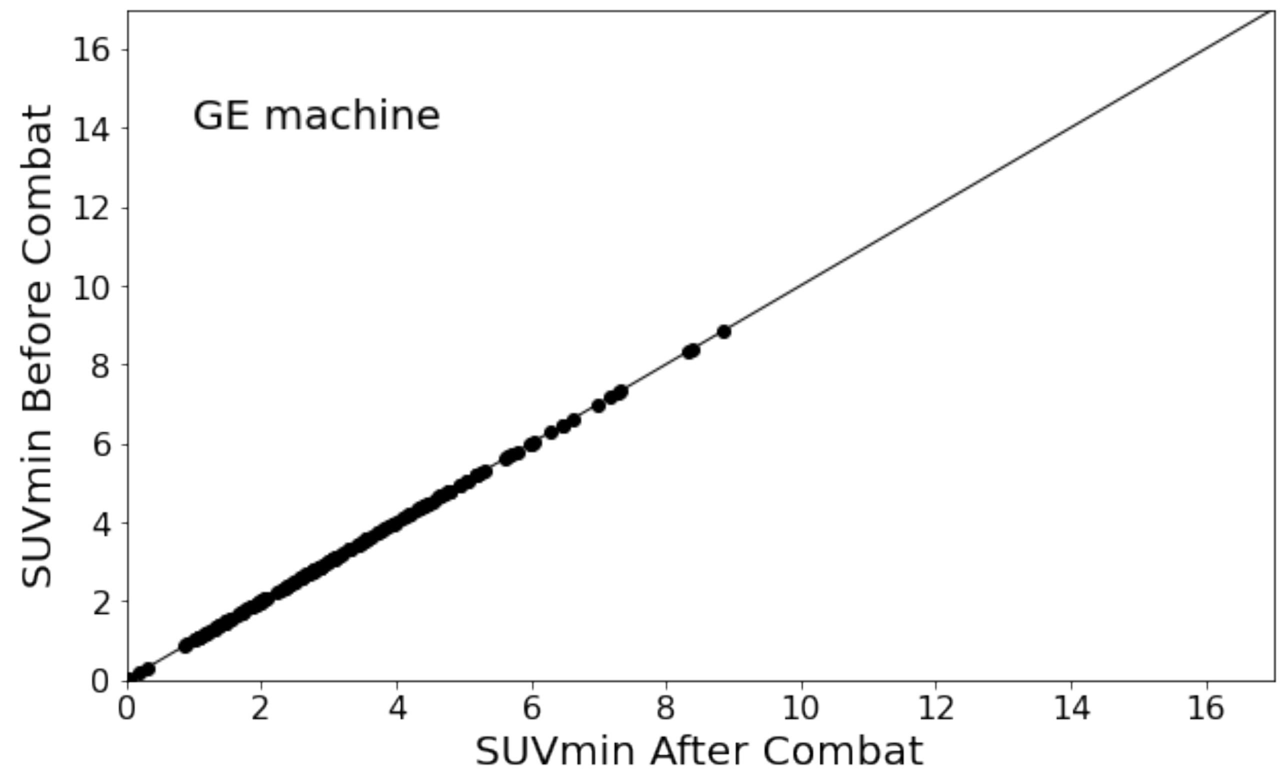
[Nioche et al. Cancer Research 2018]

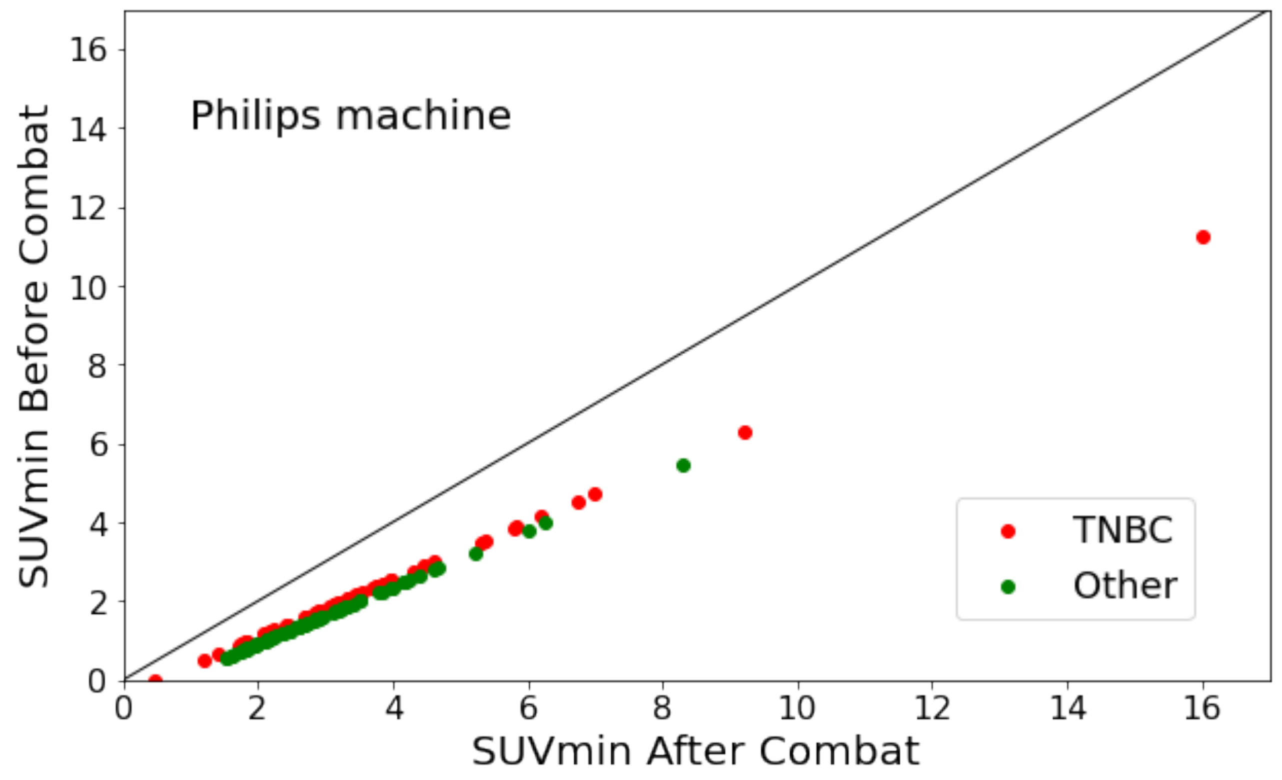
Data **harmonization**

- We use the **neuroCombat** function (Python library) to perform multi-scanner harmonization of the data
- 2 scanners: GE and Philips
- We harmonize the radiomic features
- We specify a biological covariate: cancer type (TNBC or Other)
- We use the GE scanner data as the reference batch for harmonization

Triple-negative breast cancer

LUMinal: hormone-receptor positive, HER2 negative and has low levels of Ki-67
HER
LUM-HER



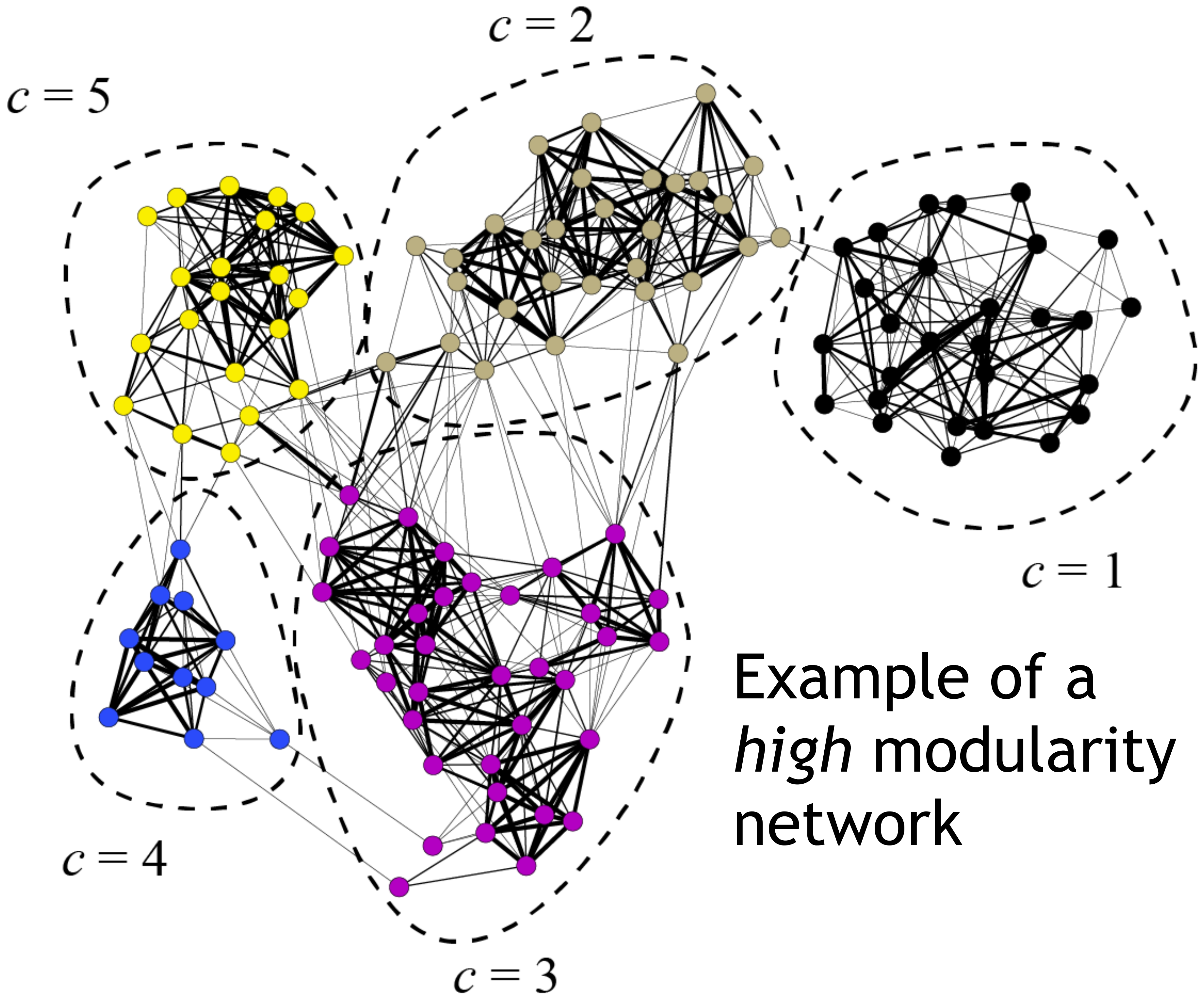


Results using the *Tumor ROI* radiomics

[Orlhac et al. *A post-reconstruction harmonization method for multicenter radiomic studies in PET* JNM 2018]

Unsupervised clustering

- Patients are clustered using the graph-based community detection method PhenoGraph (for Python3)
- The data is represented as a network which connects phenotypically similar (Jaccard similarity metric) radiomic profiles
- Communities are extracted by optimising the network modularity, which measures the strength of division of a network into clusters (Louvain method)



Example of a *high* modularity network

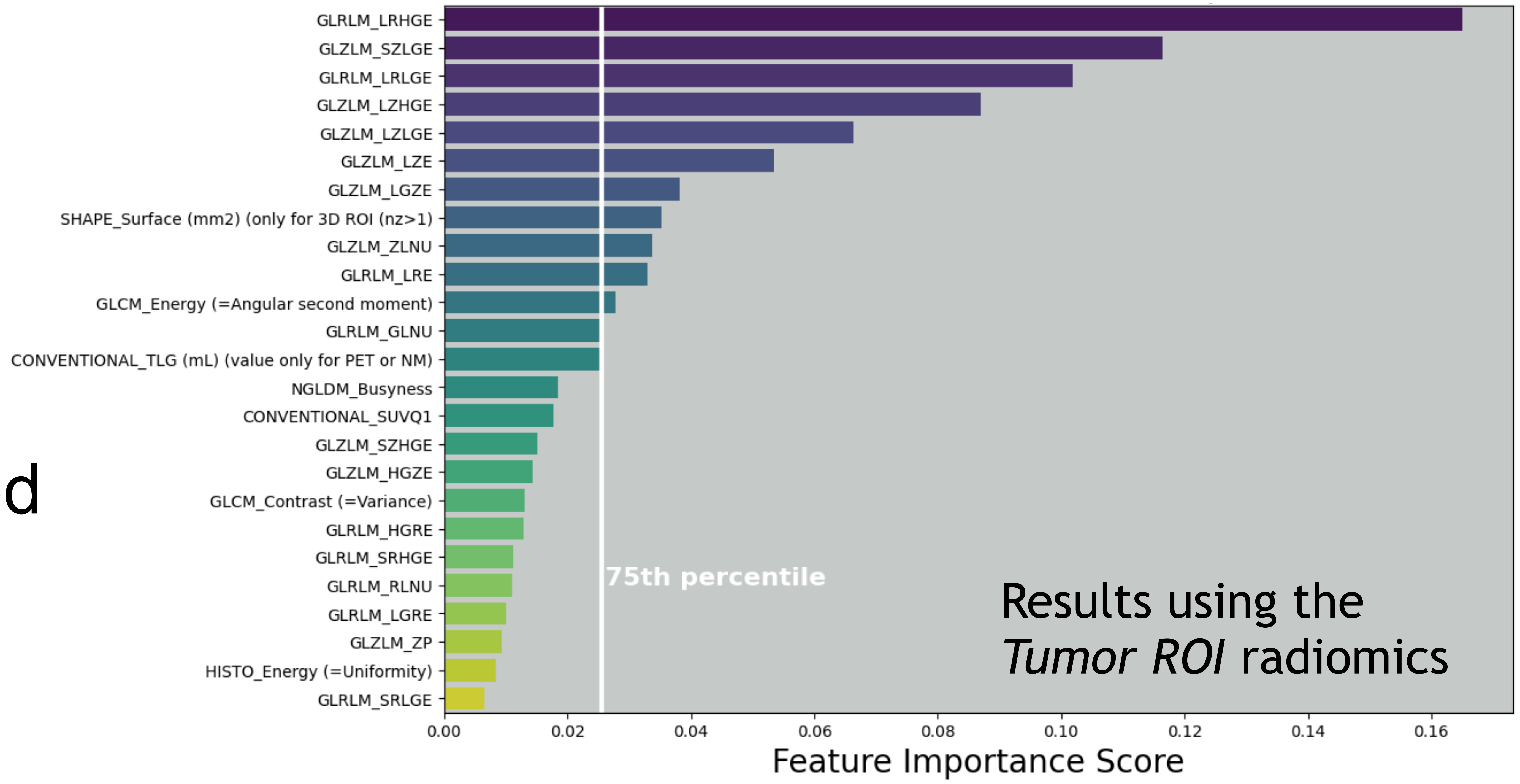
[PhenoGraph: Levine et al. Cell 2015]

[Louvain method: Blondel et al. Journal of Statistical Mechanics 2008]

Supervised extraction of important features

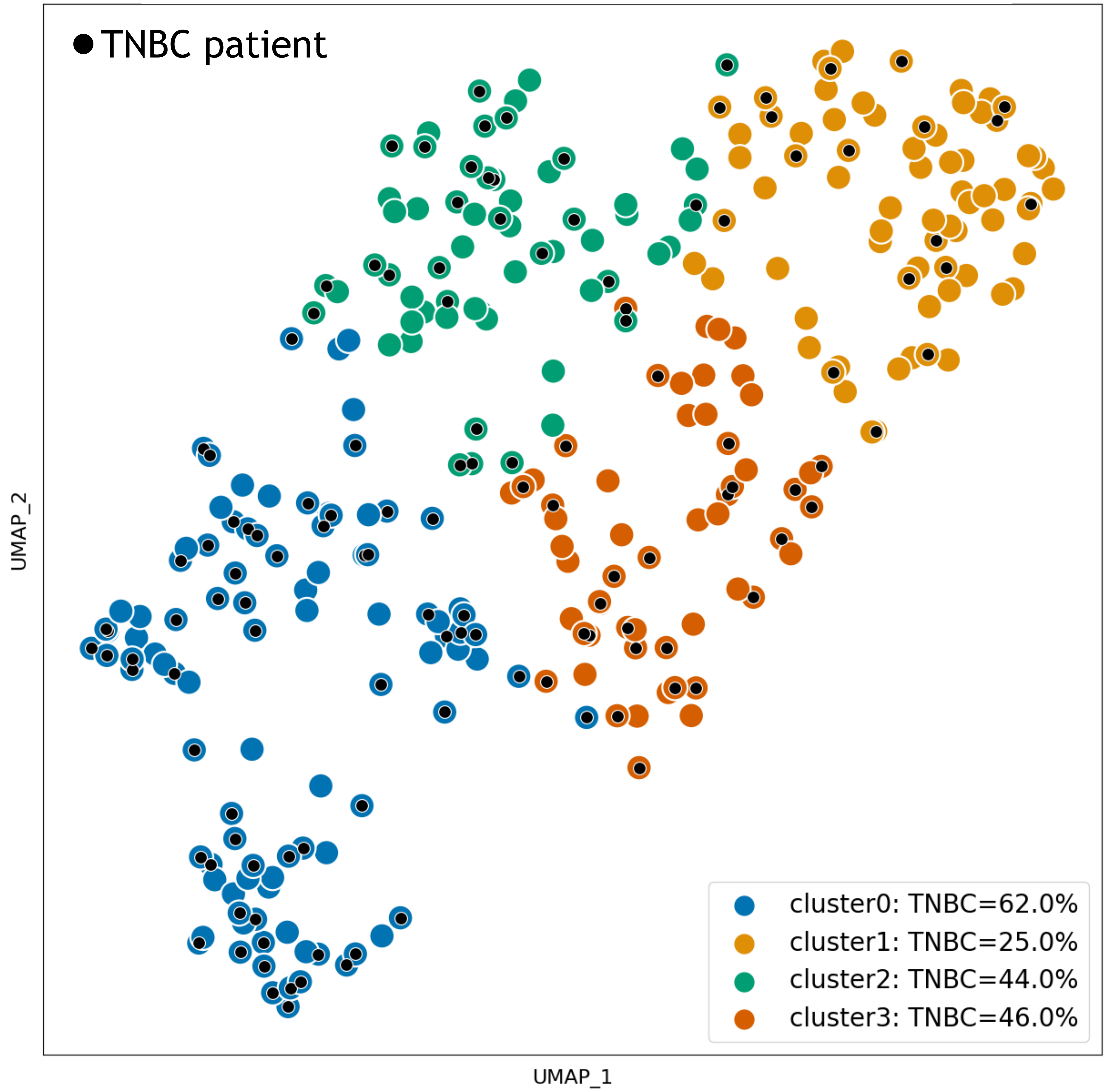
- The input data to PhenoGraph is either composed of all features or of a sub-group of features
- Features are selected using the importance scores of an optimised random forest classifier trained to predict the cancer type (TNBC or Other: LUM, HER and LUM-HER)
- Sub-groups of features are composed of features for which the importance score is greater than the 70th to 85th percentile of the scores

Example of a sub-group of features: importance score > 75th percentile

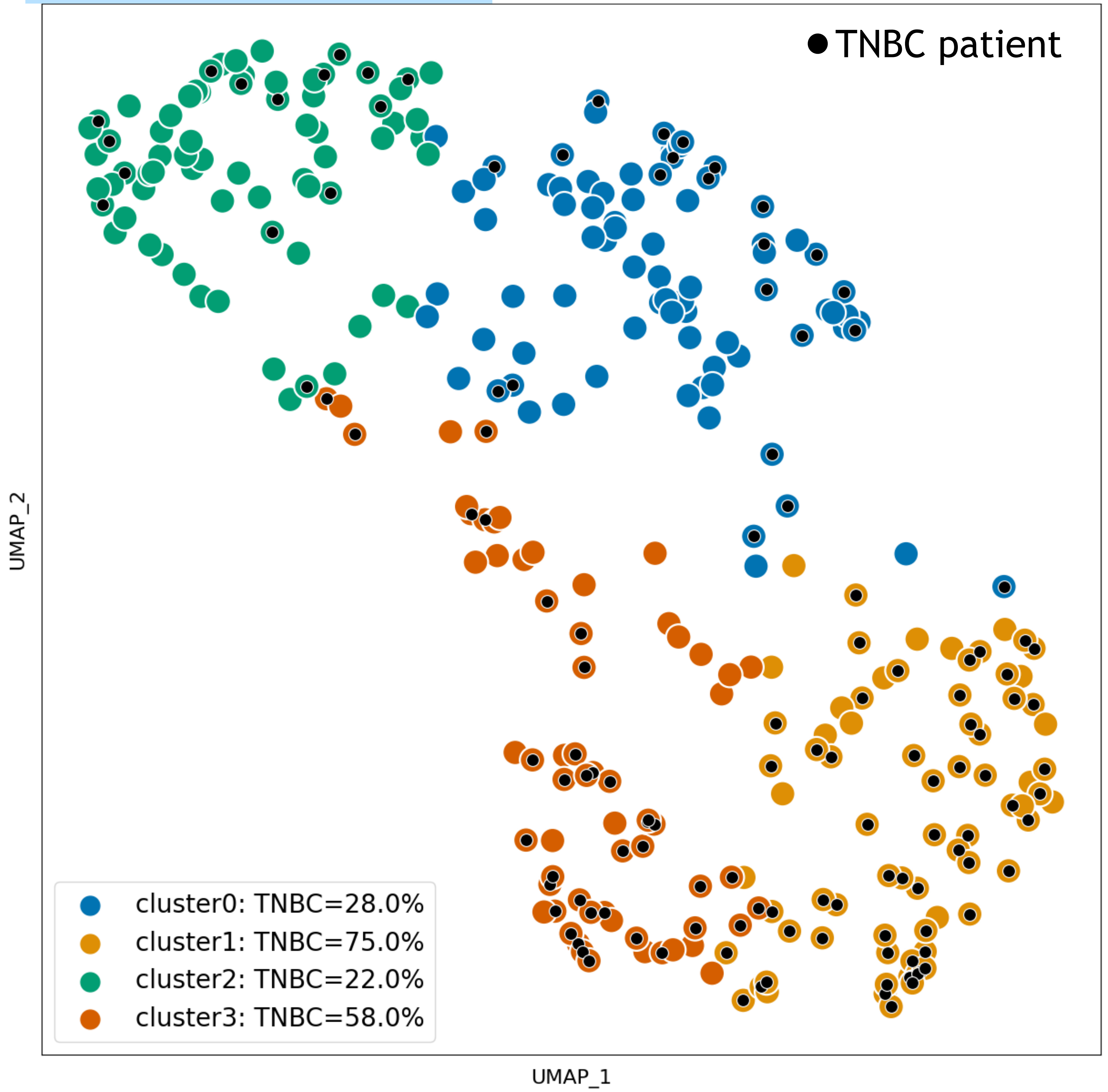


Clusters **composition** in cancer type

All features: Clusters composition in TNBC type



75th percentile features: Clusters composition in TNBC type



Results using the *Tumor ROI* radiomics

Is the repartition of patients in the clusters coherent with the available knowledge on the data, i.e. the cancer type (TNBC or Other) ?

Purity or quality of the clustering method

Forestier et al. define the **clustering purity:** [Forestier et al. KSEM 2010]

$$\boxed{\Pi} = \frac{1}{N}\sum_{i}^{K} c_i \boxed{\pi(c_i)} \quad \text{with} \quad \pi(c_i) = \sum_{j}^{C} \boxed{(\frac{n_j^i}{c_i})^2}$$

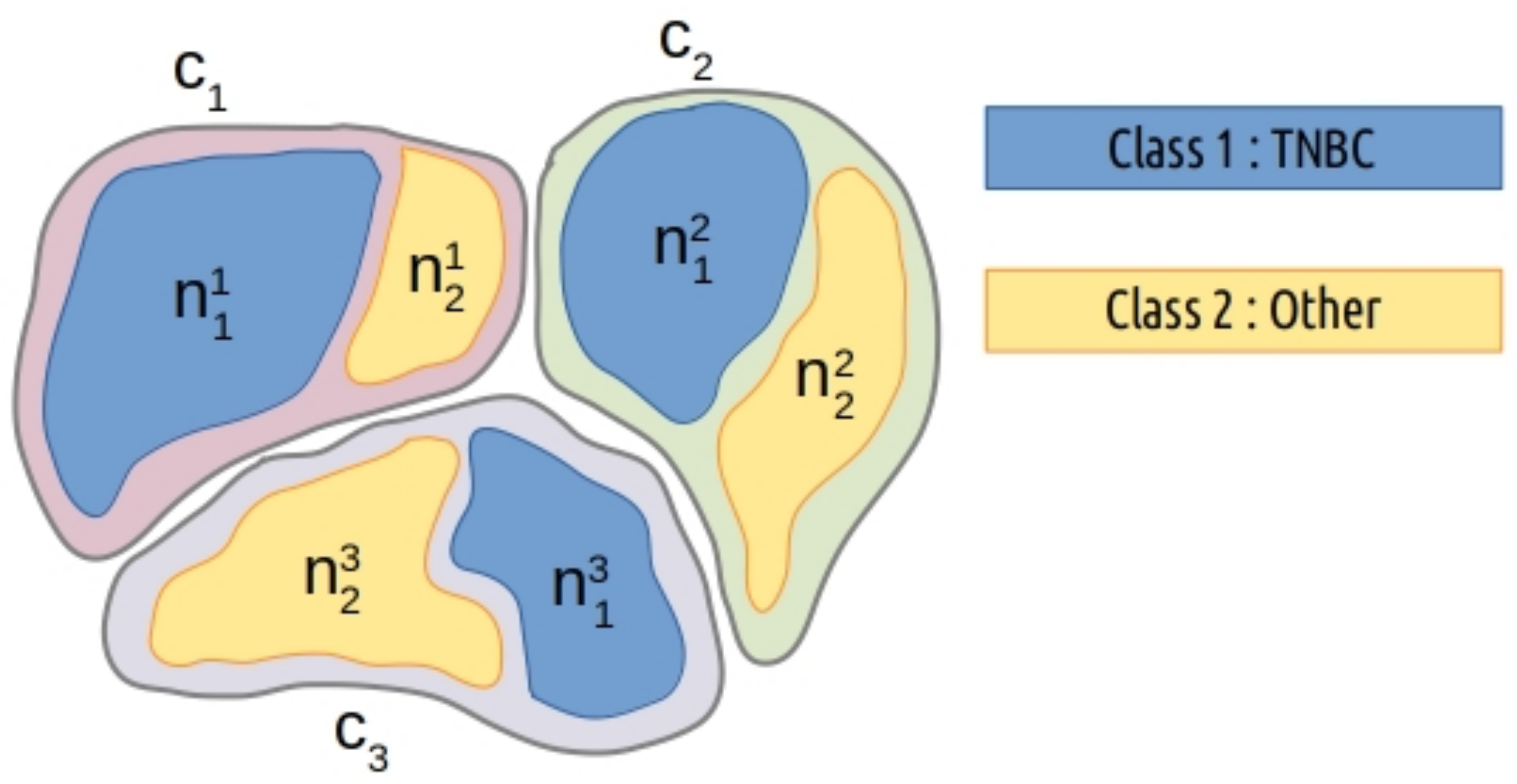
cluster's purity

Probability that, given a cluster i and 2 randomly chosen labeled patients of this cluster, they both are of the same class j

K = number of clusters

C = number of classes (in this study C=2)

c_i = number of patients in cluster i



n_j^i = number of patients of class j in cluster i

Purity or quality of the clustering method

Forestier et al. define the **clustering purity:** [Forestier et al. KSEM 2010]

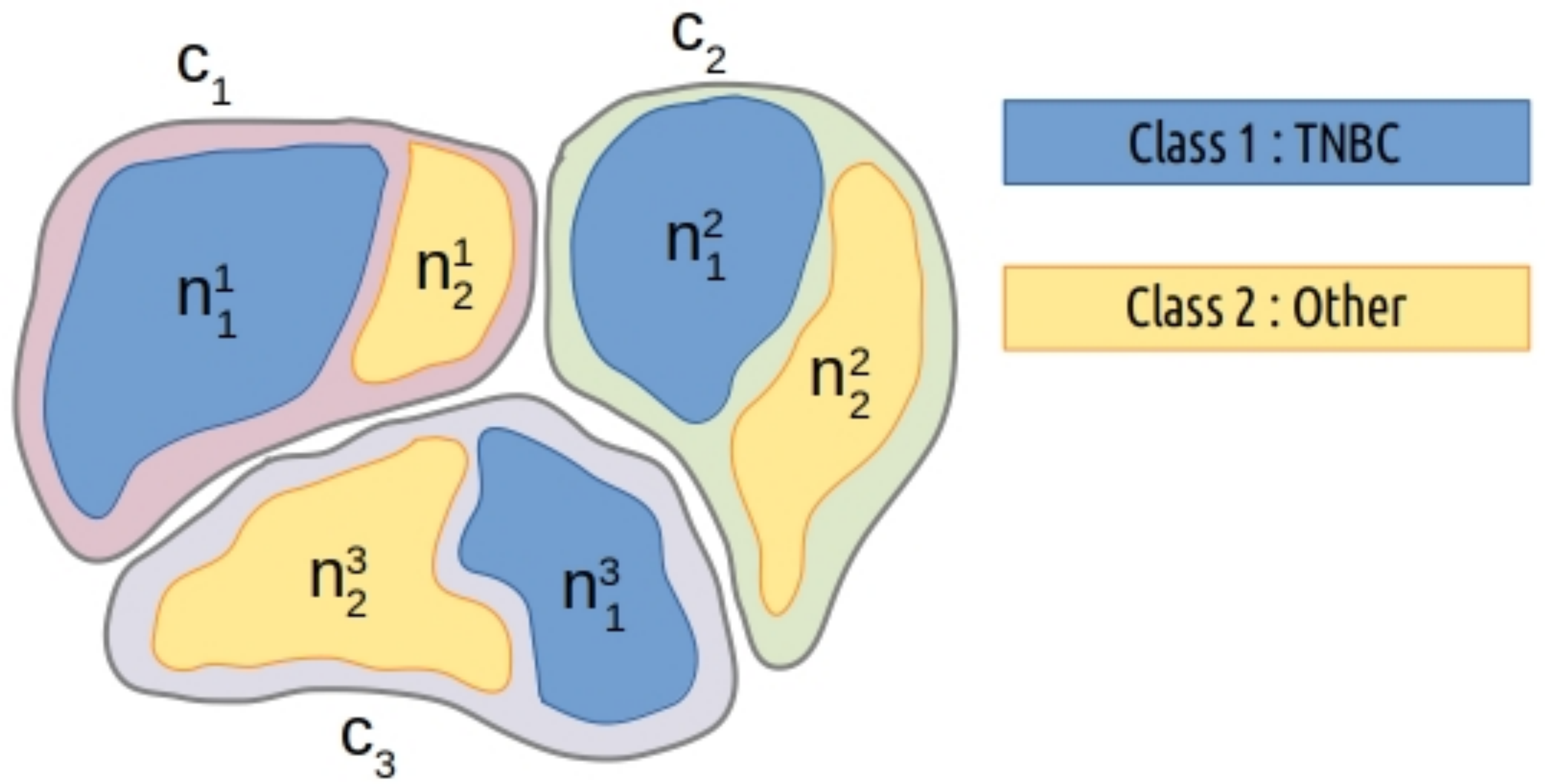
$$\boxed{\Pi} = \frac{1}{N}\sum_{i}^{K} c_i \boxed{\pi(c_i)} \quad \text{with} \quad \pi(c_i) = \sum_{j}^{C} \boxed{\left(\frac{n_j^i}{c_i}\right)^2}$$

cluster's purity

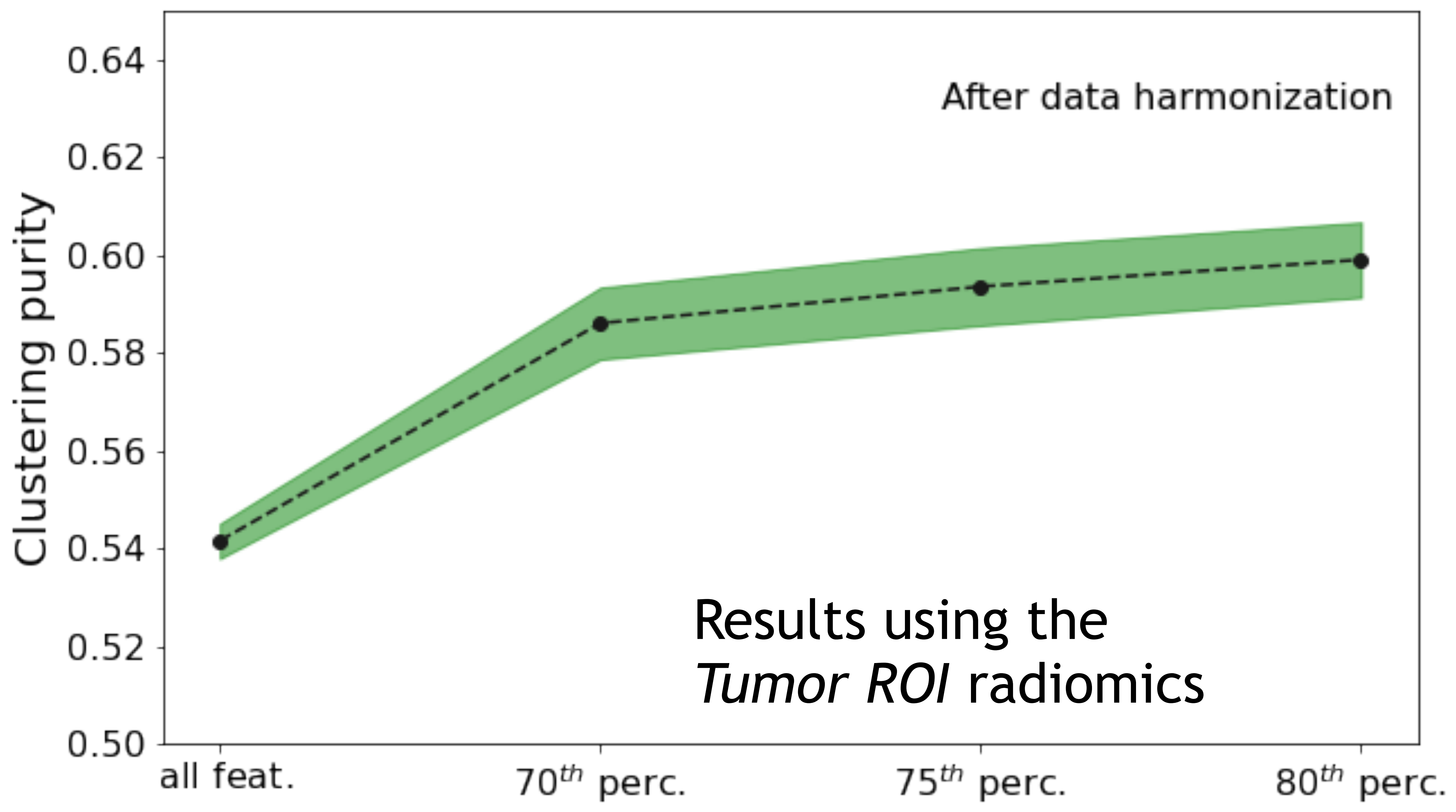
K = number of clusters

C = number of classes (in this study C=2)

c_i = number of patients in cluster i

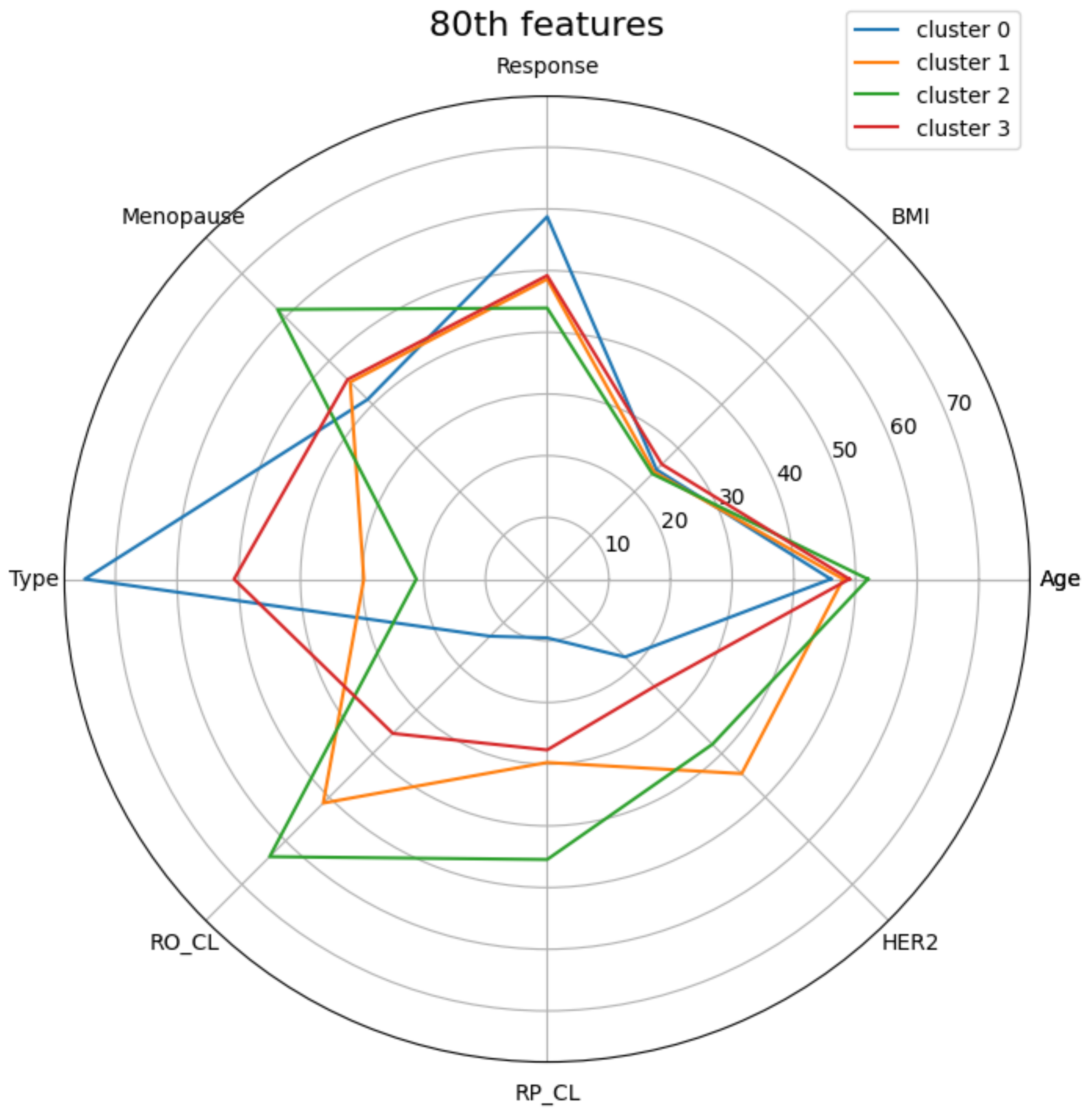


n_j^i = number of patients of class j in cluster i



Using a sub-group of important features allows for an **increase** in the clusters purity in terms of cancer types.

Comparing clusters using **radar plots**



Treatment response	1: PCR	0: NonPCR
Cancer type	1: TNBC	0: Other
Menopause status	1: Yes	0: No
RO_CL (Estrogen receptor)	1: RO+	0:RO-
RP_CL (Progesterone receptor)	1: RP+	0: RP-

PCR = Pathological Complete Response

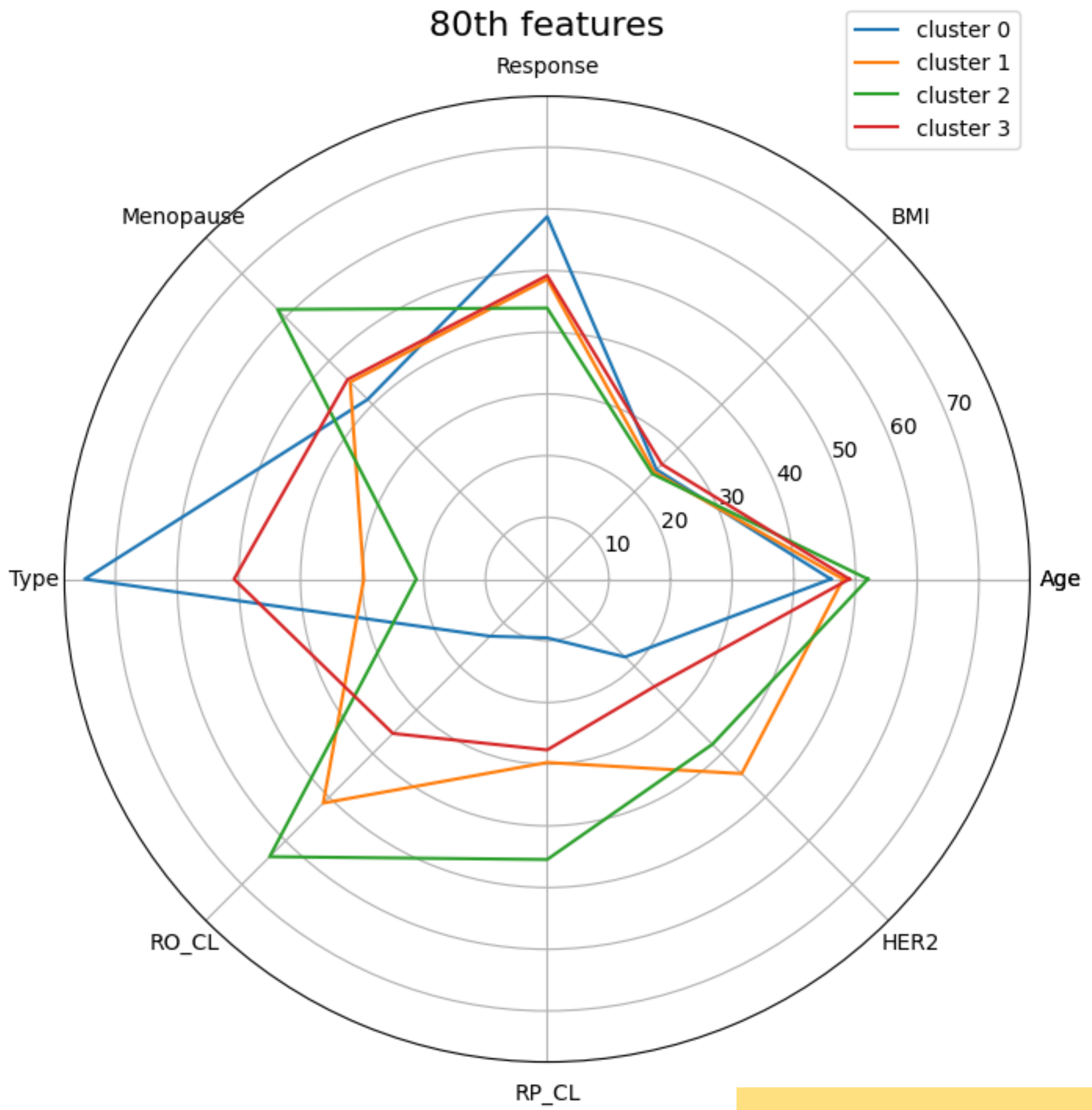
- Apart from Age and BMI, each variable is scaled in [0, 100].
- Values in the radar plot correspond to the mean value of each variable in the cluster.

What do we learn?

cluster 0: Younger patients with low hormonal receptors are mostly TNBC patients with higher rates of PCR.

cluster 2: Older patients with higher rates of hormonal receptors are mostly non-TNBC patients and have the lowest rate of PCR.

Comparing clusters using **radar plots**



Treatment response	1: PCR	0: NonPCR
Cancer type	1: TNBC	0: Other
Menopause status	1: Yes	0: No
RO_CL (Estrogen receptor)	1: RO+	0:RO-
RP_CL (Progesterone receptor)	1: RP+	0: RP-

PCR = Pathological Complete Response

- Apart from Age and BMI, each variable is scaled in [0, 100].
- Values in the radar plot correspond to the mean value of each variable in the cluster.

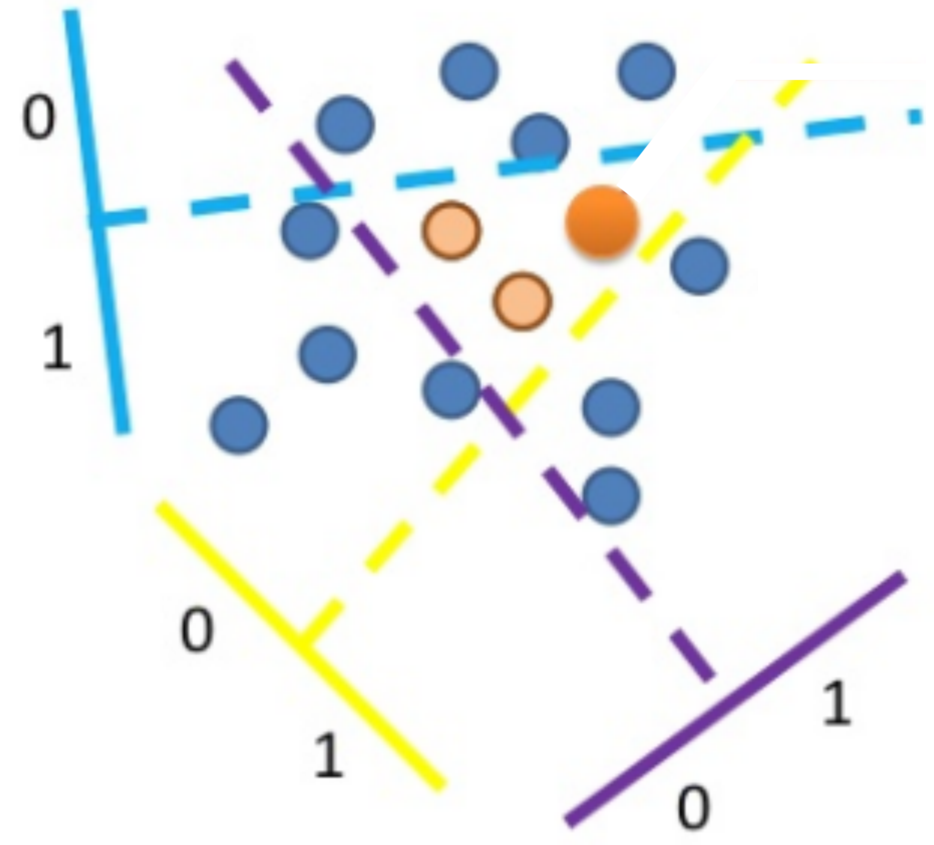
<u>What do we learn?</u>

cluster 0: Younger patients with low hormonal receptors are mostly TNBC patients with higher rates of PCR.

cluster 2: Older patients with higher rates of hormonal receptors are mostly non-TNBC patients and have the lowest rate of PCR.

Clusters obtained **from radiomics** capture **clinical characteristics** of the patients.

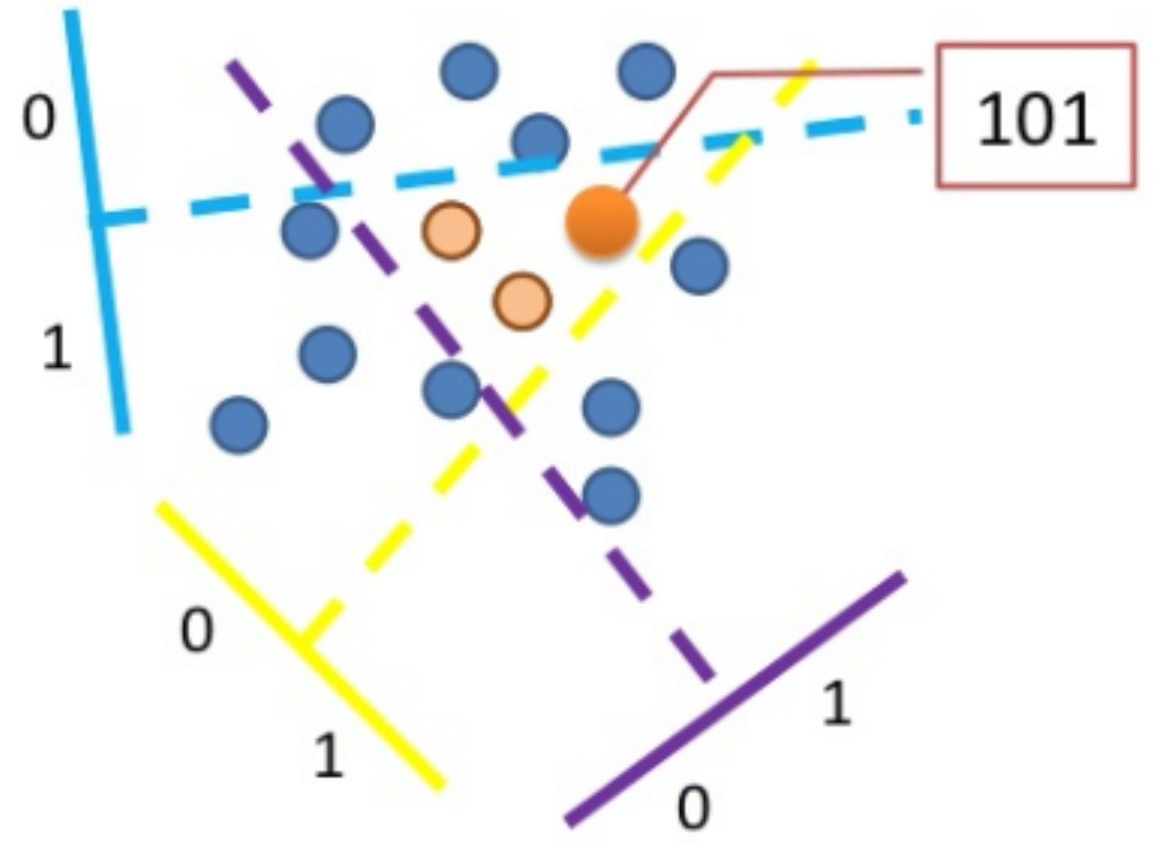
Finding nearest neighbours (**similar** patients)



- **Locality Sensitive Hashing** (LSH) is an algorithm that hashes similar items into same buckets with high probability. Since similar items end up in same buckets, this technique can be used for ***approximate* nearest neighbour search**.
- LSH partition the data into bins by **randomly drawing N hyper-planes** (of dimension = number of features).
  - How bad can this be? The chance to split 2 close points with a random hyper-plane is small. **Good performance**.

Finding nearest neighbours (**similar** patients)

- **Locality Sensitive Hashing** (LSH) is an algorithm that hashes similar items into same buckets with high probability. Since similar items end up in same buckets, this technique can be used for ***approximate* nearest neighbour search**.
- LSH partition the data into bins by **randomly drawing N hyper-planes** (of dimension = number of features).
  - How bad can this be? The chance to split 2 close points with a random hyper-plane is small. **Good performance**.

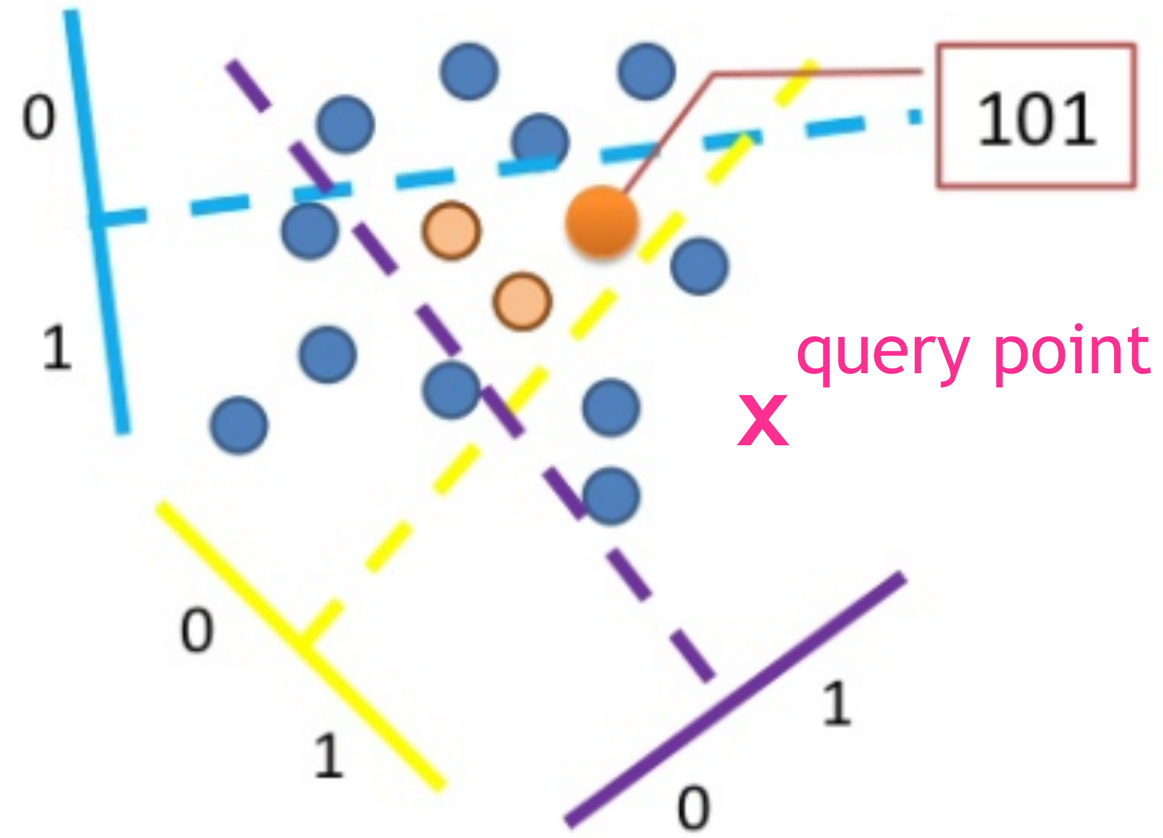


- Compute a **score** for each data point under each hyper-plane, translated into a **binary index**.
- We use a **N-bit binary vector** per data point as a bin index. The more bits two indexes have in common, the more similar their input data was.
- A **hash table** is created (one time cost to create): a table that associates the LSH bin index to a list of data points.

N-bit binary vector	[001….101]	[101….100]	[111….001]	…	[101….000]
Data points indices	{1, …, 170}	{201, …, 375}	{21, …, 410}	…	{45, …, 341}

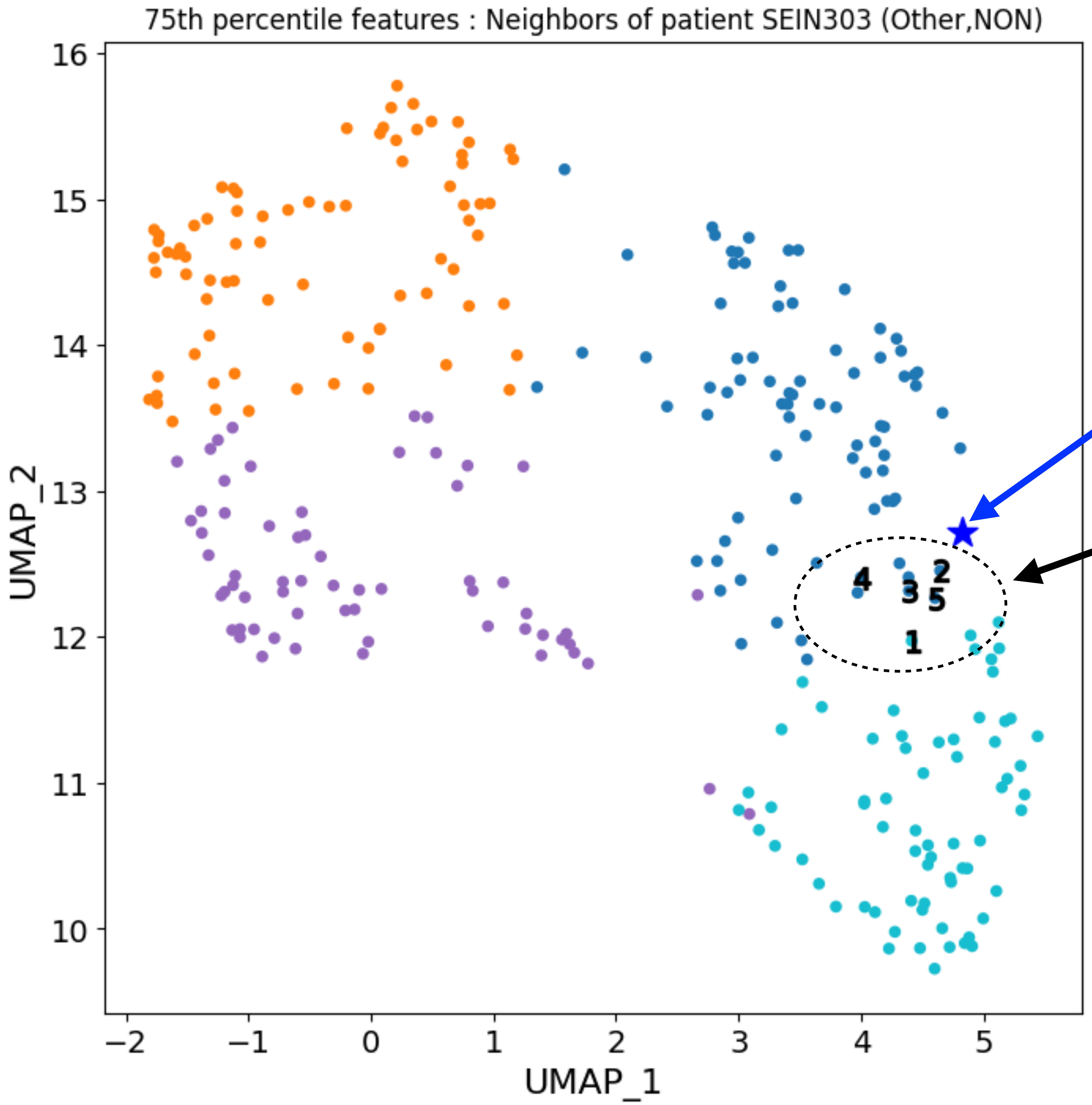
Finding nearest neighbours (**similar** patients)

- **Locality Sensitive Hashing** (LSH) is an algorithm that hashes similar items into same buckets with high probability. Since similar items end up in same buckets, this technique can be used for ***approximate* nearest neighbour search**.
- LSH partition the data into bins by **randomly drawing N hyper-planes** (of dimension = number of features).
  - How bad can this be? The chance to split 2 close points with a random hyper-plane is small. **Good performance**.



- Compute a **score** for each data point under each hyper-plane, translated into a **binary index**.
- We use a **N-bit binary vector** per data point as a bin index. The more bits two indexes have in common, the more similar their input data was.
- A **hash table** is created (one time cost to create): a table that associates the LSH bin index to a list of data points.
- We can do many **queries** on that hash table. We retrieve the data points that are hashed into the same bucket as the query point.

Finding nearest neighbours (similar patients)



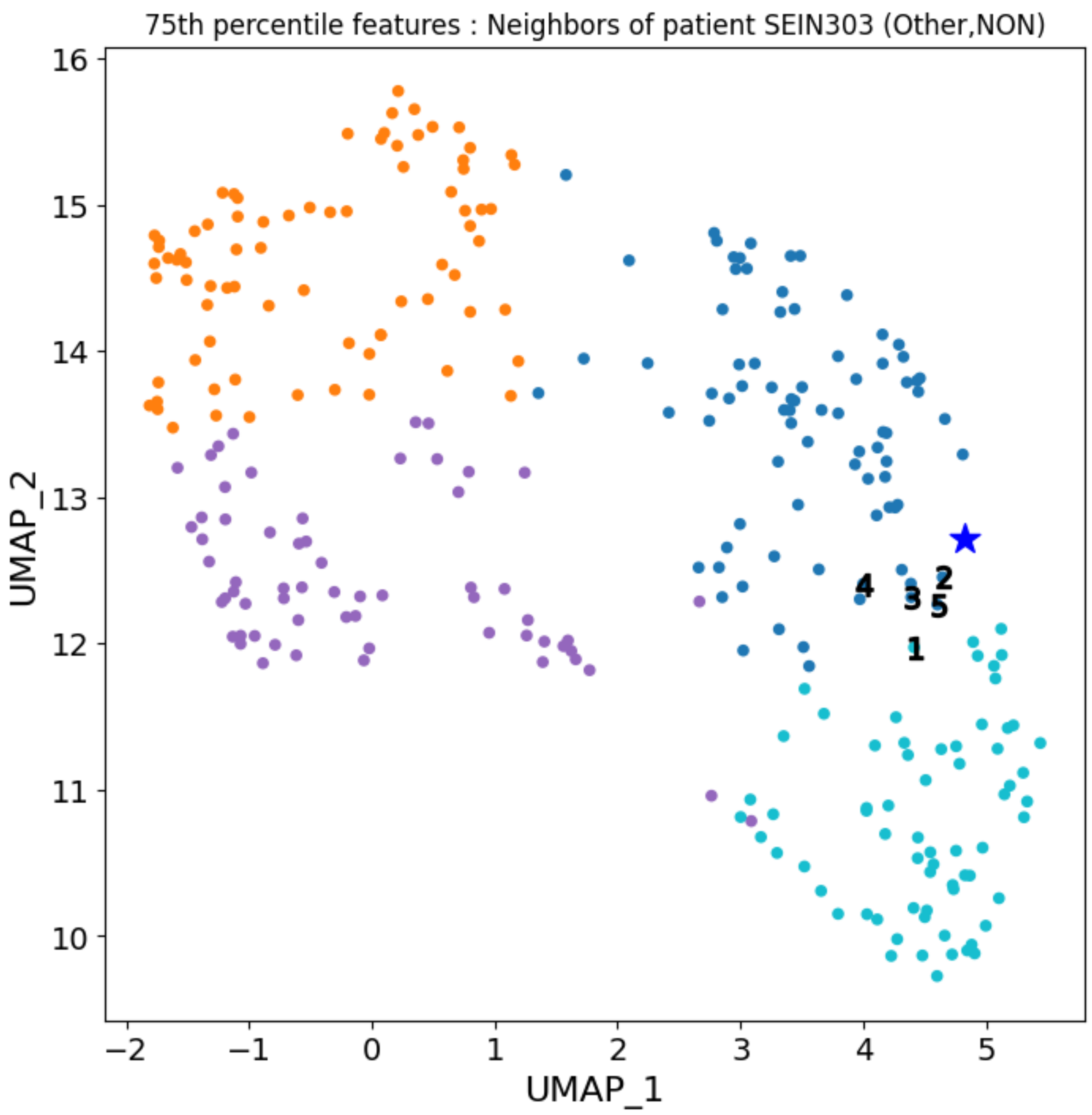
New patient (SEIN303) is projected into the clustered database.

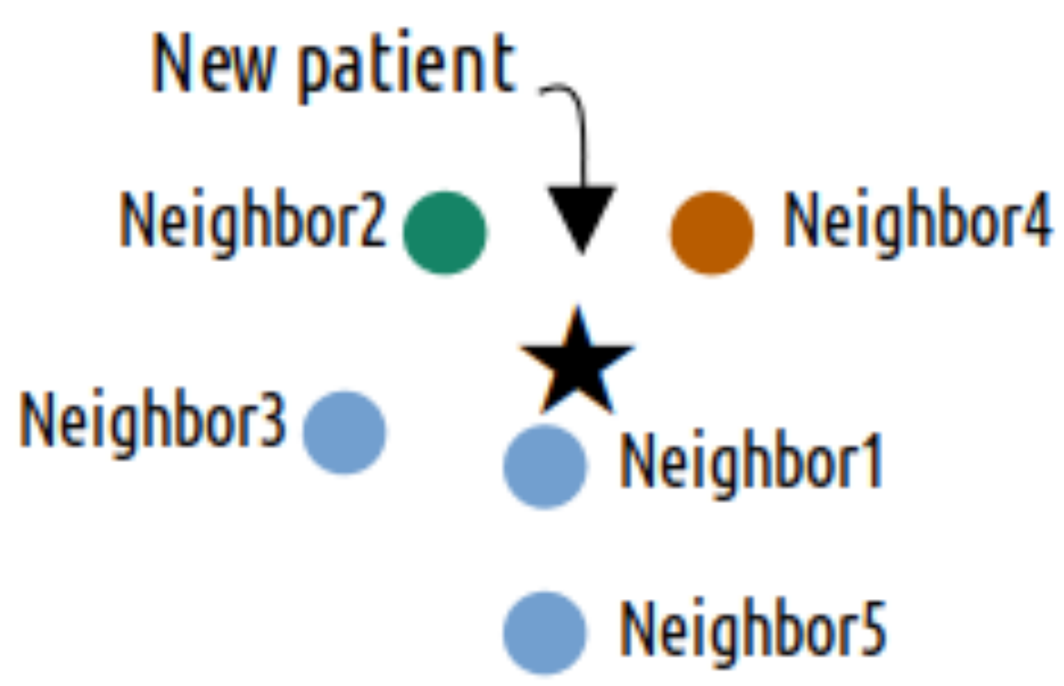
5 closest (most similar) patients obtained using the LSH algorithm.

Reminder: PANACEE main goal

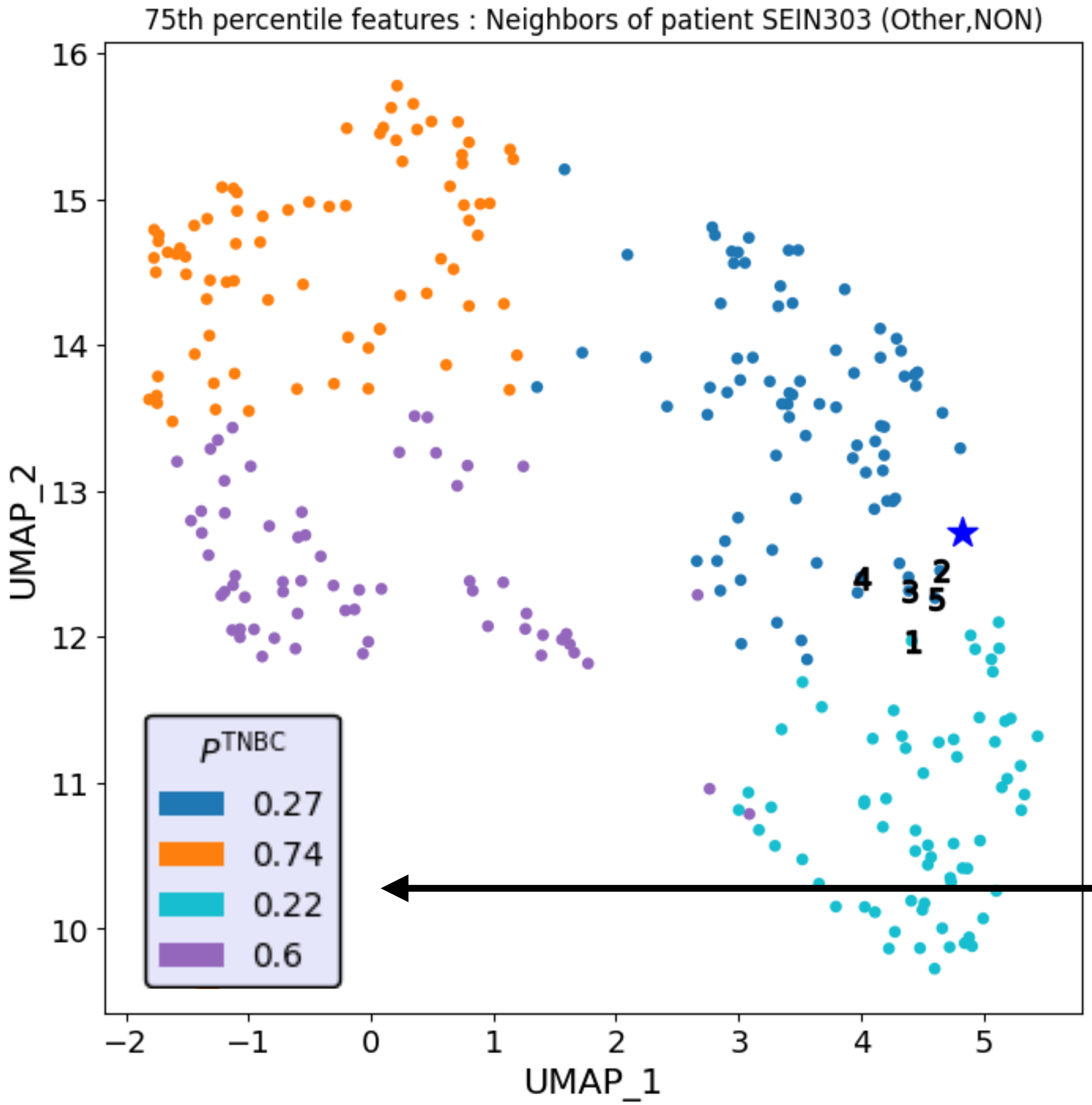
The medical history of these “twin-patients” could allow doctors to suggest the therapeutic strategy to be adopted for the new patient?

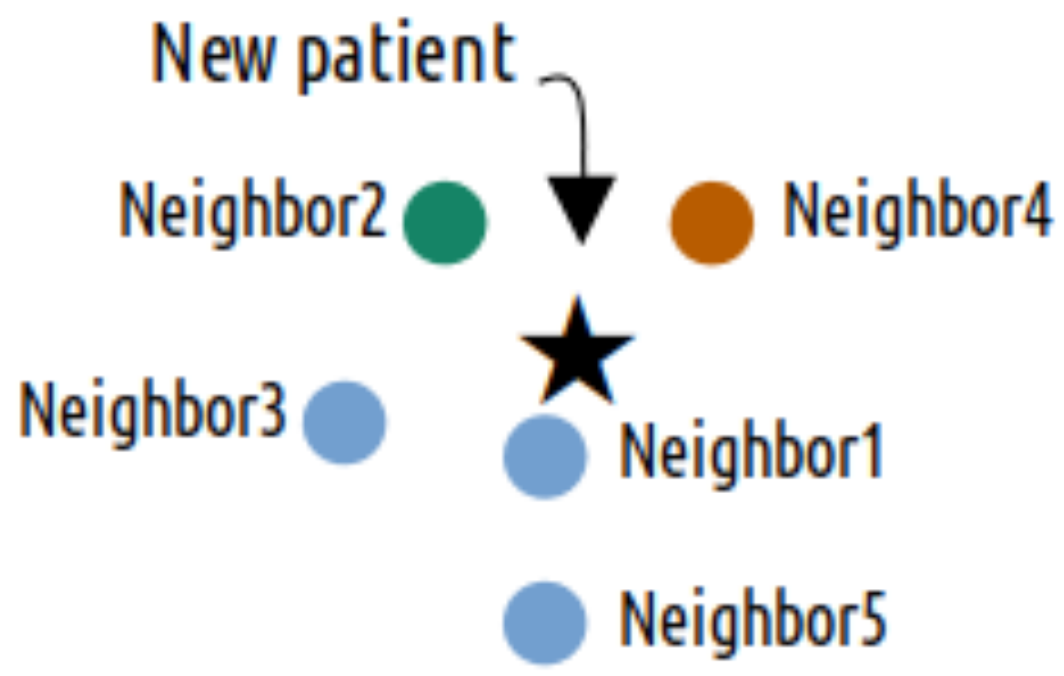
Deriving the new patient's cancer type from "twins"?





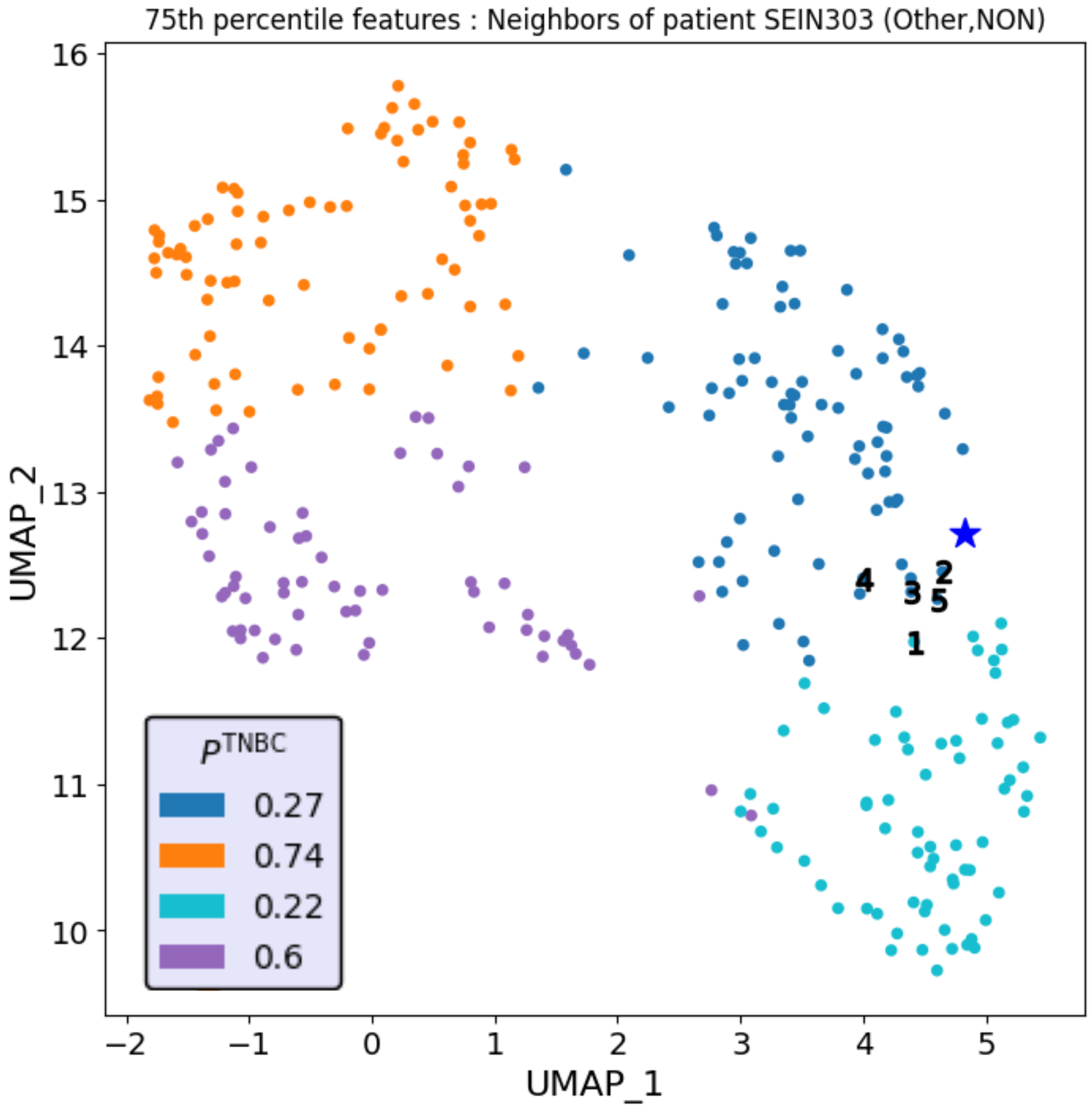
Deriving the new patient's cancer type from "twins"?

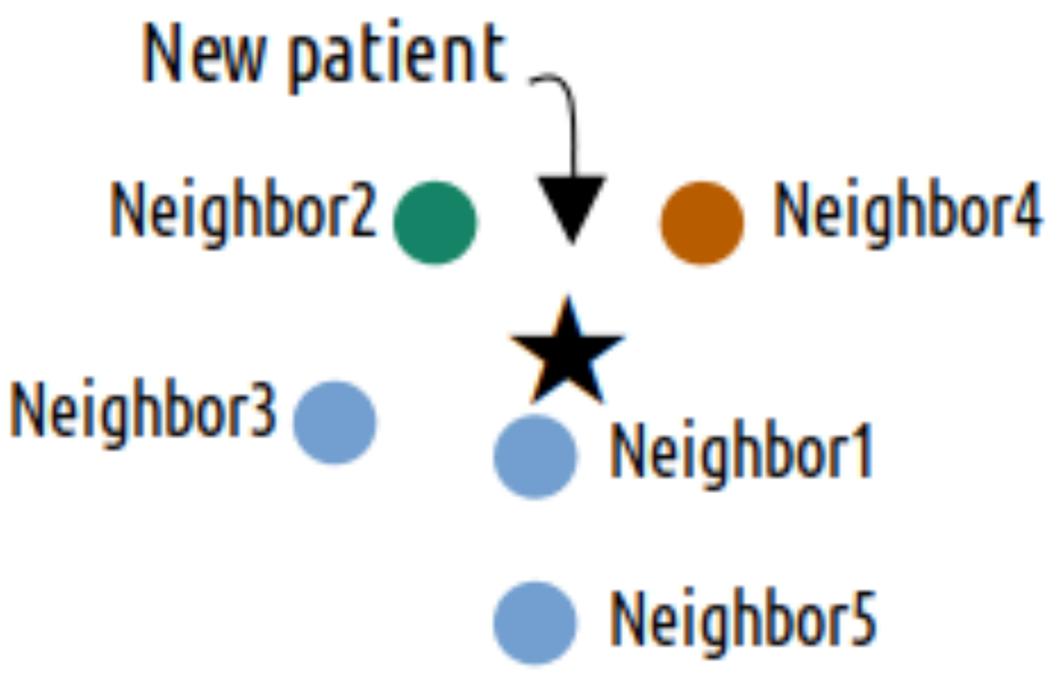




<u>Idea:</u> Use the information obtained from the PhenoGraph clustering of the RALUCA-Breast database to assign to each neighbour a probability of being TNBC.

Deriving the new patient's cancer type from "twins"?

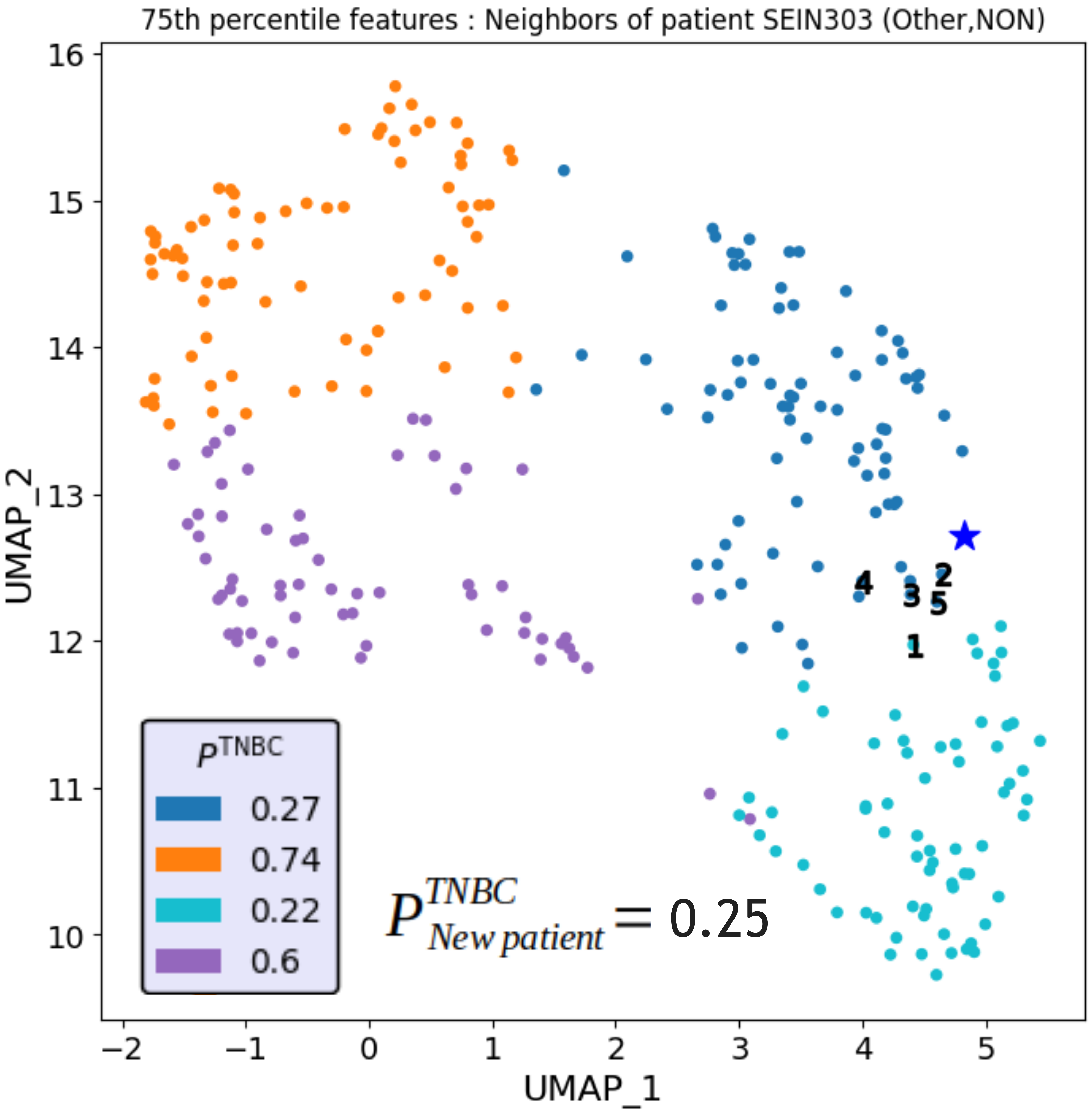


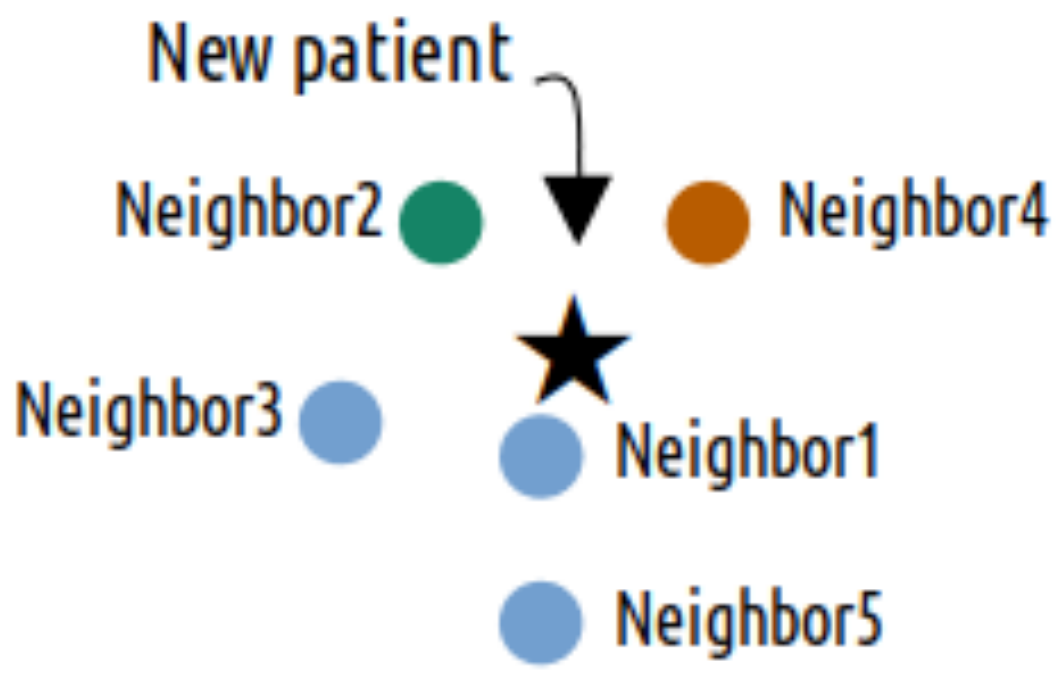


$$P^{TNBC}_{New\ patient} = \frac{\sum_{n=1}^{5} \frac{p^{TNBC}_{cluster\ number\ of\ Neighbor\ n}}{d_{Neighbor\ n}}}{\sum_{n=1}^{5} \frac{1}{d_{Neighbor\ n}}}$$

Mean probability including a weighting factor that takes into account the distance to the nearest neighbour.

Deriving the new patient's cancer type from "twins"?

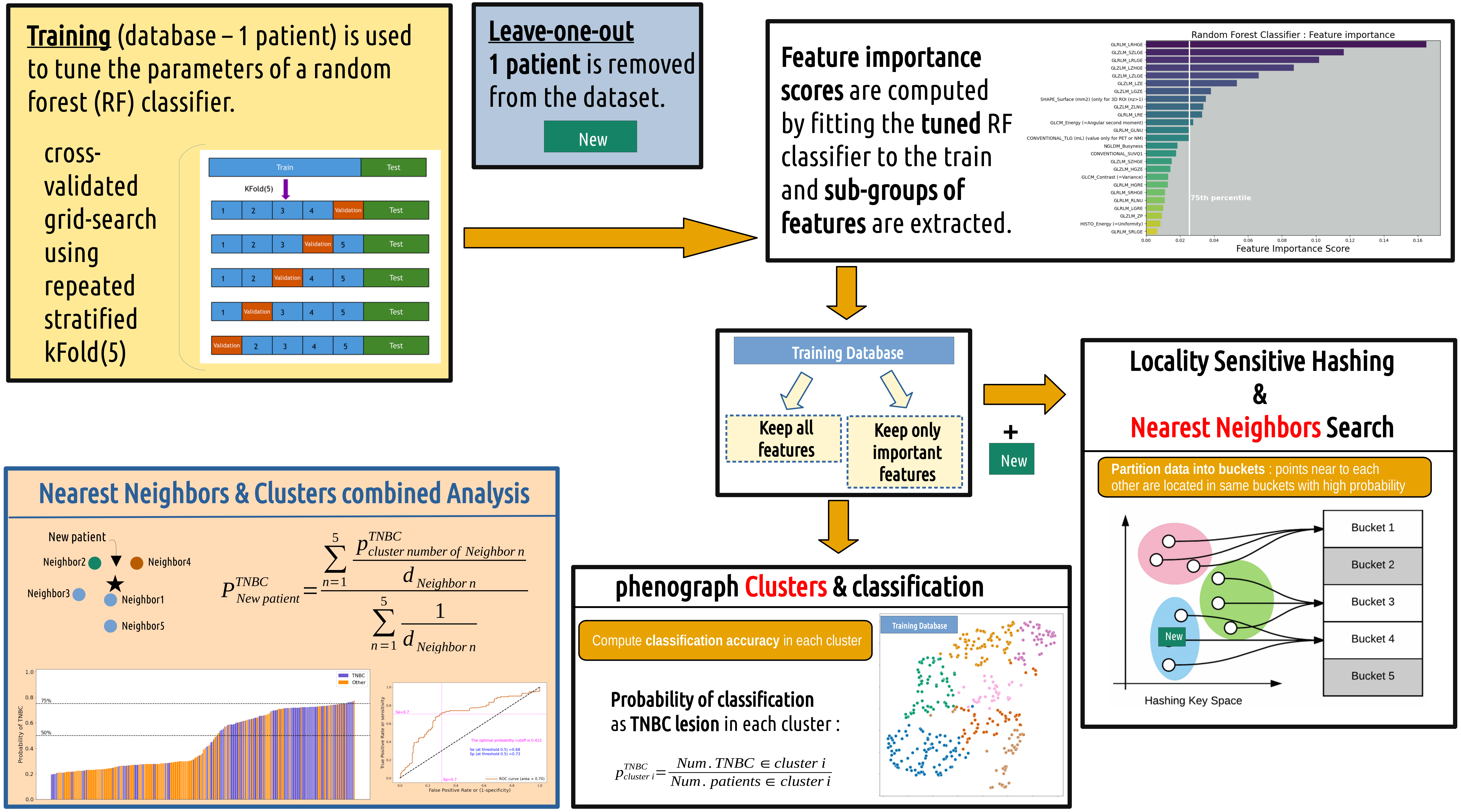




$$P^{TNBC}_{New\,patient} = \frac{\sum_{n=1}^{5} \frac{P^{TNBC}_{cluster\ number\ of\ Neighbor\,n}}{d_{Neighbor\,n}}}{\sum_{n=1}^{5} \frac{1}{d_{Neighbor\,n}}}$$

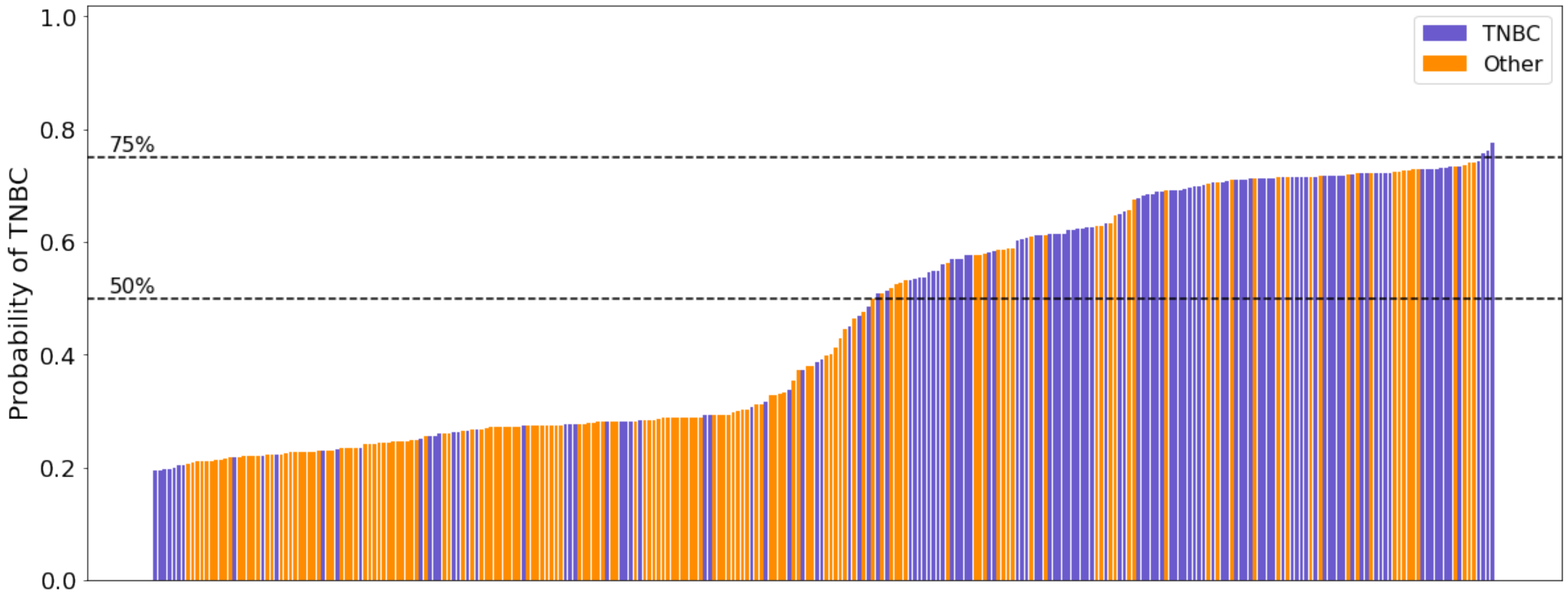
Mean probability including a weighting factor that takes into account the distance to the nearest neighbour.

Cancer type classification performance

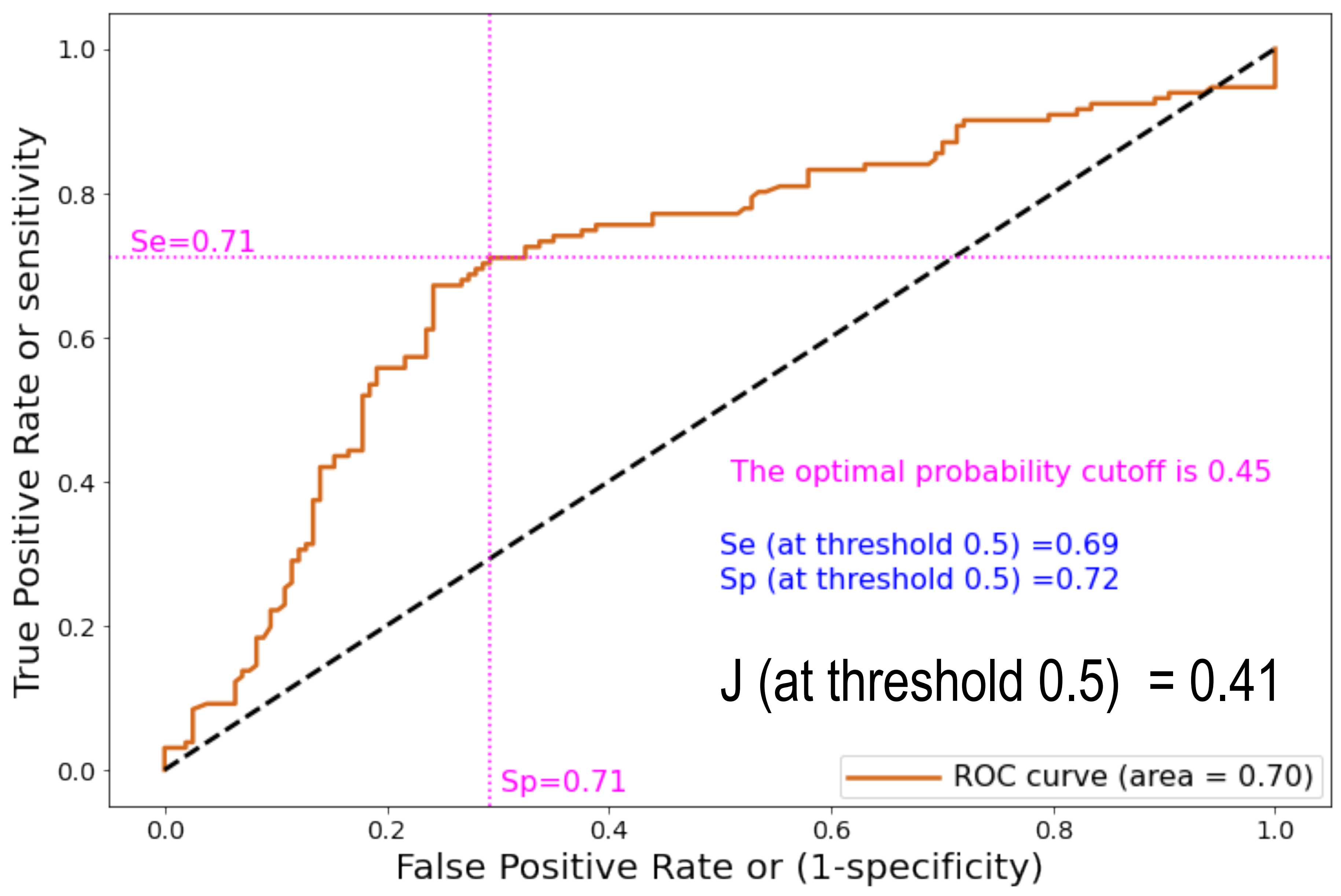


Cancer type classification performance

Tumor radiomics and **75th** percentile features

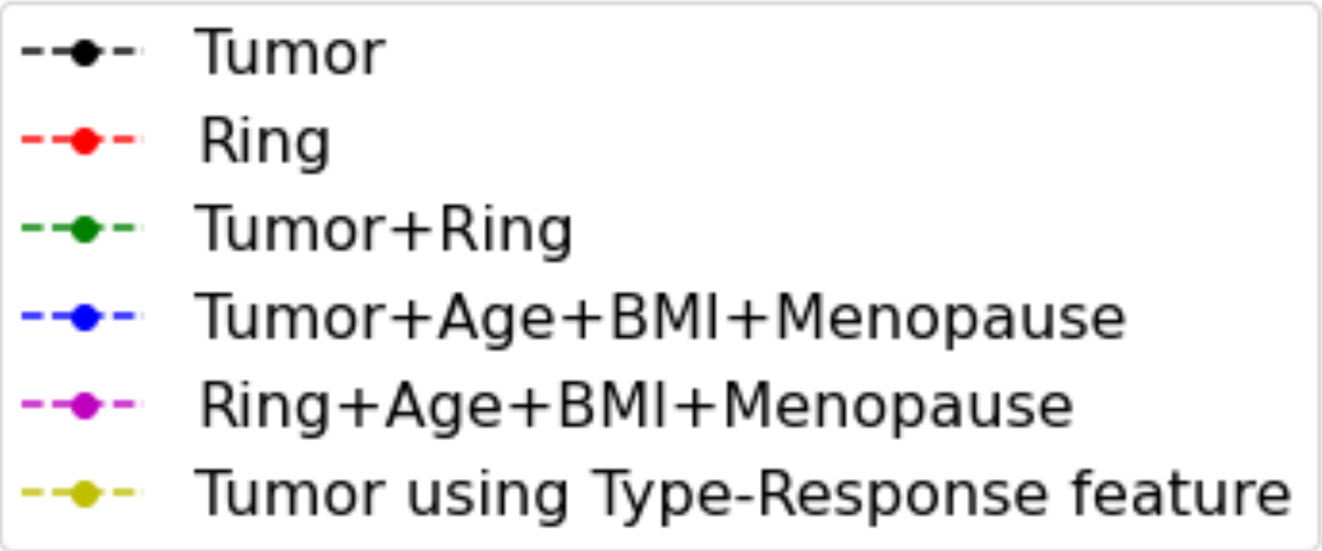


Mostly **Other** patients here.... Mostly **TNBC** patients here....



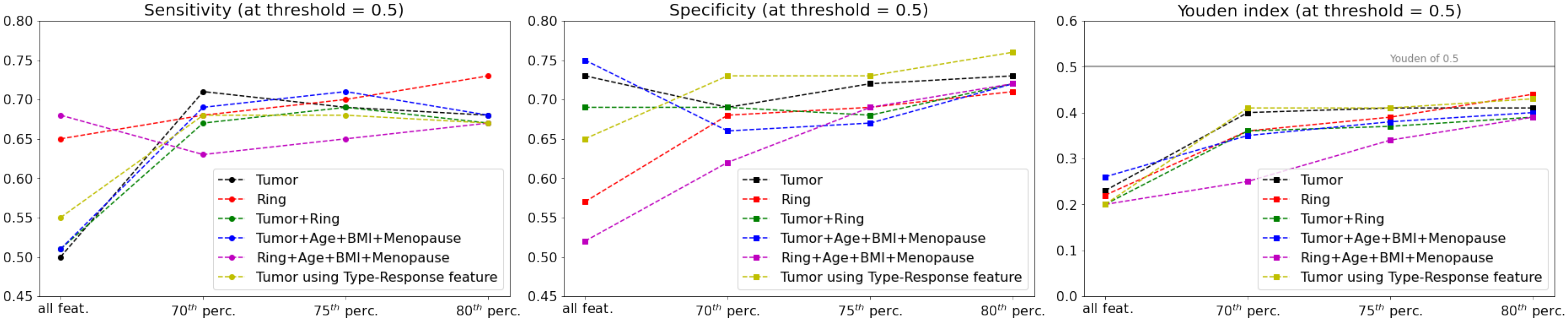
The optimal cutoff would be where the sensitivity and specificity are high.

Cancer type classification performances: other scenarios



Classification using radiomics and clinical features from different VOI

Classification using Tumor radiomics and a new feature (association of cancer type and treatment response: 4 states) as the target of the random forest classifier (used for features extraction)



Best results (highest Youden index) are obtained with these 2 scenarios:

- 80^{th} percentile sub-group of features (cancer type is used as the target to extract important features) and the Ring VOI
- 80^{th} percentile sub-group of features (type-response is used as the target to extract important features) and the Tumor VOI

Conclusion

- We propose a **semi-supervised** (un-supervised clustering + supervised features extraction) method to find **similarities** between patients from a database.
- Our findings are:
  - Using a **sub-group of important features increases** the clustering **purity**.
  - Un-supervised clusters obtained **from radiomics** capture **clinical characteristics**.
  - Applying this method on RALUCA-Breast (289 patients) shows **good performances** in classifying the cancer type (TNBC versus Other).
- Additional findings (not discussed in the presentation):
  - Unfortunately when trying to predict the **treatment outcome** (PCR or Non-PCR) for patients with TNBC breast cancer, the performances are not good: AUC ~ 0.5
    - We think that this prediction is rather complex for breast cancer
    - Maybe the prediction is less complex for lung cancer patients? (to do list)

Perspectives

- **Increase** the RALUCA-Lung database (so far 58 patients were segmented and the segmentations were reviewed by M. Luporsi)
- But, in total we only have clinical informations for 79 patients, so the lung DB will still be small at the end
- **Continue** working with **RALUCA-Breast** data (289 patients): **Predict** cancer type from neighbours using **alternative methods** and compare classifier performances:
    - First neighbour type
    - True types from the 5 closest neighbours
    - Majority vote among the 5 closest neighbours
    - Looking at *all* neighbours within a **distance** from the new patient; define that distance by looking at the distributions of all distances between patients and the distances to the first neighbour.