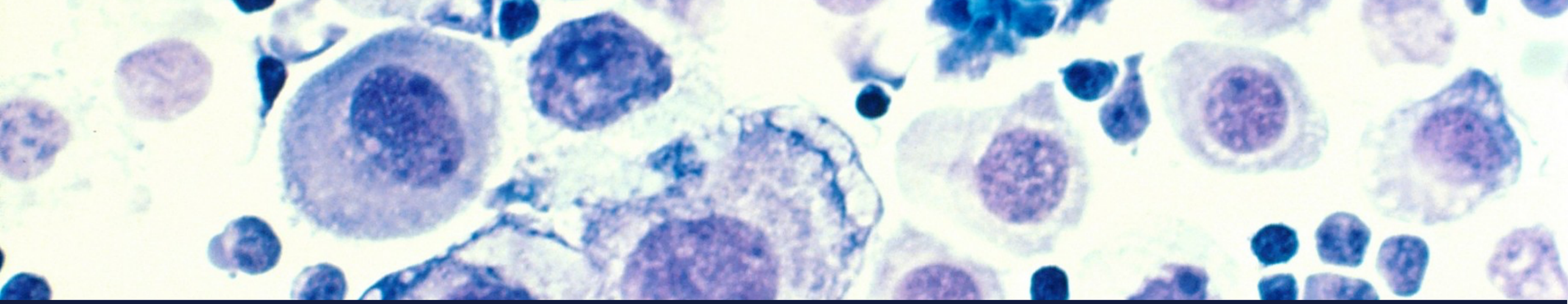
PROFILING BREAST CANCER CELLS THROUGH PCA

LUDOVICO GANDOLFI
EDOARDO FERRERO



13%

of women will develop a breast carcinoma in their life

30%

of all cancers diagnosed to women

15%

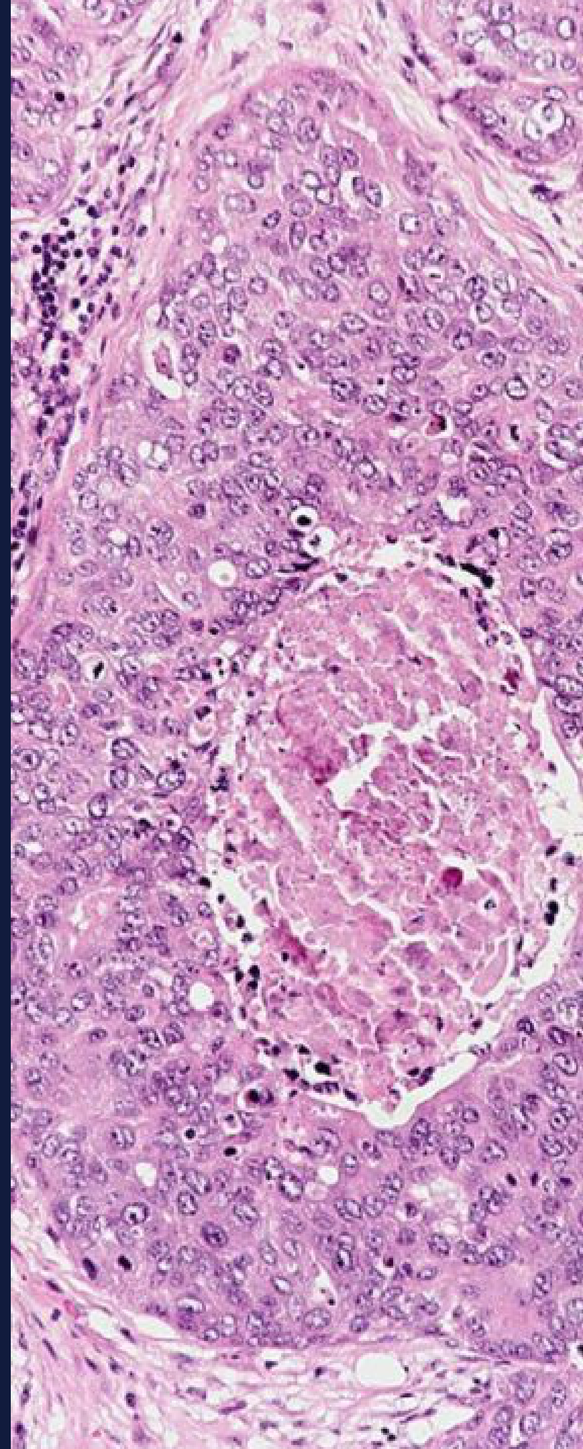
of diagnosed breast cancers are supposed to be inheritors

| Data retrieved from: https://www.breastcancer.org/facts-statistics |

Breast Carcinoma

Normally, the cells in our bodies replace themselves through an orderly process of cell growth, but over time, mutations can “turn on” certain genes and “turn off” others in a cell. That changed cell gains the ability to keep dividing without control or order, producing more cells just like it and forming a tumor.

| Data retrieved from: https://www.breastcancer.org/about-breast-cancer |



Which cellular features of a malignant breast cancer could be aggregated?

Our research question

Dataset

Breast Cancer Wisconsin Data Set
- University of Wisconsin

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe 30 characteristics of the cell nuclei derived from the 568 digitalized images.

id	oncavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	per
842302	0,3001	0,1471	0,2419	0,07871	1.095	0,9053	
842517	0,0869	0,07017	0,1812	0,05667	0,5435	0,7339	
84300903	0,1974	0,1279	0,2069	0,05999	0,7456	0,7869	
84348301	0,2414	0,1052	0,2597	0,09744	0,4956	1.156	
84358402	0,198	0,1043	0,1809	0,05883	0,7572	0,7813	
843786	0,1578	0,08089	0,2087	0,07613	0,3345	0,8902	
844359	0,1127	0,074	0,1794	0,05742	0,4467	0,7732	
84458202	0,09366	0,05985	0,2196	0,07451	0,5835	1.377	
844981	0,1859	0,09353	0,235	0,07389	0,3063	1.002	
84501001	0,2273	0,08543	0,203	0,08243	0,2976	1.599	
845636	0,03299	0,03323	0,1528	0,05697	0,3795	1.187	
84610002	0,09954	0,06606	0,1842	0,06082	0,5058	0,9849	
846226	0,2065	0,1118	0,2397	0,078	0,9555	3.568	
846381	0,09938	0,05364	0,1847	0,05338	0,4033	1.078	
84667401	0,2128	0,08025	0,2069	0,07682	0,2121	1.169	
84799002	0,1639	0,07364	0,2303	0,07077	0,37	1.033	
848406	0,07395	0,05259	0,1586	0,05922	0,4727	1,24	
84862001	0,1722	0,1028	0,2164	0,07356	0,5692	1.073	
849014	0,1479	0,09498	0,1582	0,05395	0,7582	1.017	
8510426	0,06664	0,04781	0,1885	0,05766	0,2699	0,7886	
8510653	0,04568	0,0311	0,1967	0,06811	0,1852	0,7477	
8510824	0,02956	0,02076	0,1815	0,06905	0,2773	0,9768	
8511133	0,2077	0,09756	0,2521	0,07032	0,4388	0,7096	
851509	0,1097	0,08632	0,1769	0,05278	0,6917	1.127	
852552	0,1525	0,0917	0,1995	0,0633	0,8068	0,9017	
852631	0,2229	0,1401	0,304	0,07413	1.046	0,976	
852763	0,1425	0,08783	0,2252	0,06924	0,2545	0,9832	
852781	0,149	0,07731	0,1697	0,05699	0,8529	1.849	
852973	0,1683	0,08751	0,1926	0,0654	0,439	1.012	
853201	0,09875	0,07953	0,1739	0,06149	0,6003	0,8225	
853401	0,2319	0,1244	0,2183	0,06197	0,8307	1.466	
853612	0,1218	0,05182	0,2301	0,07799	0,4825	1,03	
85382601	0,2417	0,1203	0,2248	0,06382	0,6009	1.398	
854002	0,1657	0,07593	0,1853	0,06261	0,5558	0,6062	
854039	0,1354	0,07752	0,1998	0,06515	0,334	0,6857	
854253	0,1348	0,06018	0,1896	0,05656	0,4615	0,9197	
854268	0,1319	0,05598	0,1885	0,06125	0,286	1.019	
854941	0,02562	0,02923	0,1467	0,05863	0,1839	2.342	
855133	0,02398	0,02899	0,1565	0,05504	1.214	2.188	
855138	0,1063	0,05439	0,172	0,06419	0,213	0,5914	
855167	0,0311	0,02031	0,1784	0,05587	0,2385	0,8265	
855563	0,1044	0,05669	0,1895	0,0687	0,2366	1.428	

Data cleaning

In order to investigate the peculiar characteristics of malignant carcinoma, we decided to remove from our dataset the occurrences related to cancer cells that have been classified as benign through a filtering condition on R.
The filtered dataset comprises of 212 occurrences.

id	oncavity_mean	concave points_mean	symmetry_mean	fractal_dimension_mean	radius_se	texture_se	per
842302	0,3001	0,1471	0,2419	0,07871	1.095	0,9053	
842517	0,0869	0,07017	0,1812	0,05667	0,5435	0,7339	
84300903	0,1974	0,1279	0,2069	0,05999	0,7456	0,7869	
84348301	0,2414	0,1052	0,2597	0,09744	0,4956	1.156	
84358402	0,198	0,1043	0,1809	0,05883	0,7572	0,7813	
843786	0,1578	0,08089	0,2087	0,07613	0,3345	0,8902	
844359	0,1127	0,074	0,1794	0,05742	0,4467	0,7732	
84458202	0,09366	0,05985	0,2196	0,07451	0,5835	1.377	
844981	0,1859	0,09353	0,235	0,07389	0,3063	1.002	
84501001	0,2273	0,08543	0,203	0,08243	0,2976	1.599	
845636	0,03299	0,03323	0,1528	0,05697	0,3795	1.187	
84610002	0,09954	0,06606	0,1842	0,06082	0,5058	0,9849	
846226	0,2065	0,1118	0,2397	0,078	0,9555	3.568	
846381	0,09938	0,05364	0,1847	0,05338	0,4033	1.078	
84667401	0,2128	0,08025	0,2069	0,07682	0,2121	1.169	
84799002	0,1639	0,07364	0,2303	0,07077	0,37	1.033	
848406	0,07395	0,05259	0,1586	0,05922	0,4727	1,24	
84862001	0,1722	0,1028	0,2164	0,07356	0,5692	1.073	
849014	0,1479	0,09498	0,1582	0,05395	0,7582	1.017	
8510426	0,06664	0,04781	0,1885	0,05766	0,2699	0,7886	
8510653	0,04568	0,0311	0,1967	0,06811	0,1852	0,7477	
8510824	0,02956	0,02076	0,1815	0,06905	0,2773	0,9768	
8511133	0,2077	0,09756	0,2521	0,07032	0,4388	0,7096	
851509	0,1097	0,08632	0,1769	0,05278	0,6917	1.127	
852552	0,1525	0,0917	0,1995	0,0633	0,8068	0,9017	
852631	0,2229	0,1401	0,304	0,07413	1.046	0,976	
852763	0,1425	0,08783	0,2252	0,06924	0,2545	0,9832	
852781	0,149	0,07731	0,1697	0,05699	0,8529	1.849	
852973	0,1683	0,08751	0,1926	0,0654	0,439	1.012	
853201	0,09875	0,07953	0,1739	0,06149	0,6003	0,8225	
853401	0,2319	0,1244	0,2183	0,06197	0,8307	1.466	
853612	0,1218	0,05182	0,2301	0,07799	0,4825	1,03	
85382601	0,2417	0,1203	0,2248	0,06382	0,6009	1.398	
854002	0,1657	0,07593	0,1853	0,06261	0,5558	0,6062	
854039	0,1354	0,07752	0,1998	0,06515	0,334	0,6857	
854253	0,1348	0,06018	0,1896	0,05656	0,4615	0,9197	
854268	0,1319	0,05598	0,1885	0,06125	0,286	1.019	
854941	0,02562	0,02923	0,1467	0,05863	0,1839	2.342	
855133	0,02398	0,02899	0,1565	0,05504	1.214	2.188	
855138	0,1063	0,05439	0,172	0,06419	0,213	0,5914	
855167	0,0311	0,02031	0,1784	0,05587	0,2385	0,8265	
855563	0,1044	0,05669	0,1895	0,0687	0,2366	1.428	

After conducting an exploratory analysis of the data, we conducted a Pre-Factor Analysis using the Radiant add-in on R Studio.

The analysis has been conducted on all the features related to the analyzed cell characteristics and the first results have been satisfying.

In fact, the two main performance metrics scored the following:

- The Bartlett's Test of Sphericity shows a p-value < 0.001 which displays a very strong significance.
- The Kaiser-Meyer-Olkin Test pictures a sampling adequacy for the model of 0.79 , well above the threshold set at 0.6 .

Other interesting insights can be derived from the graphical output presented in the following slide.

Pre-factor analysis diagnostics

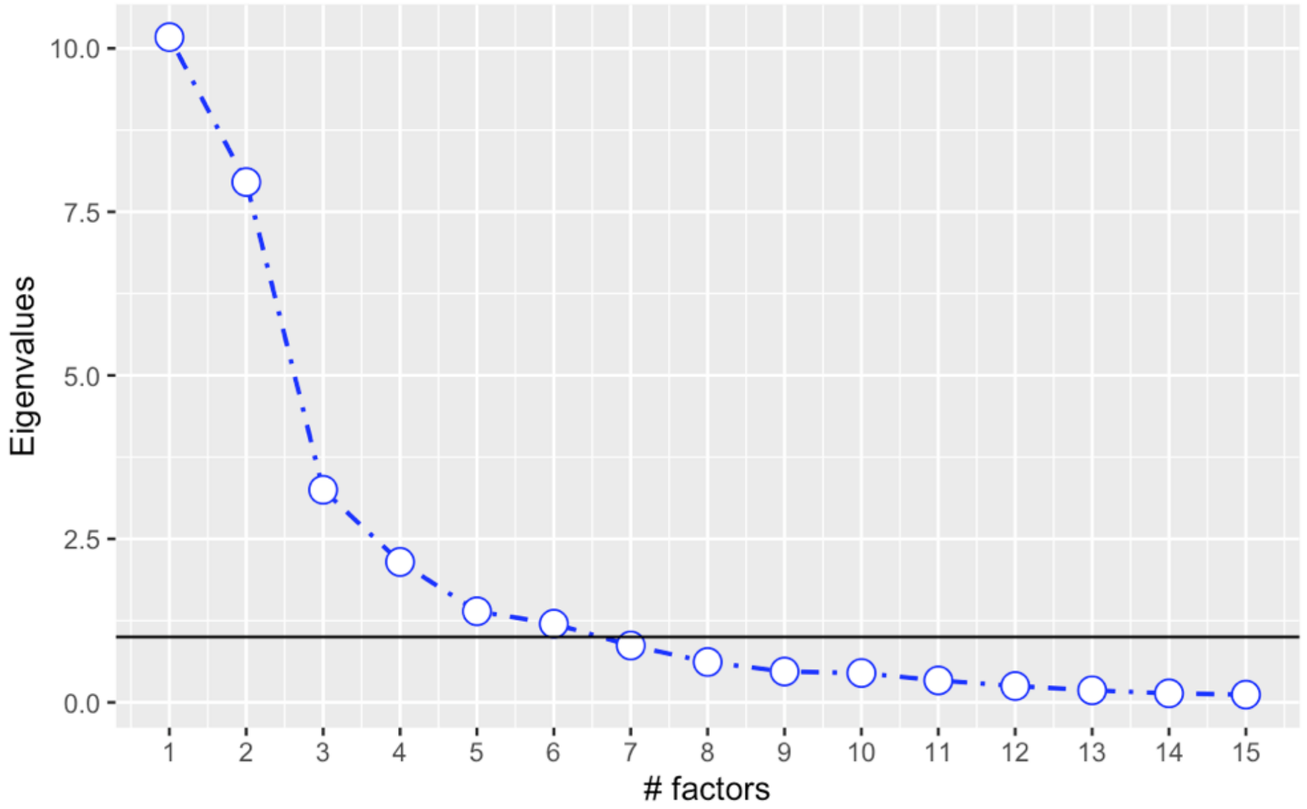
Bartlett test

Null hyp. : variables are not correlated

Alt. hyp. : variables are correlated

Chi-square: 13494.17 df(435), p.value < .001

KMO test: 0.79



Variable collinearity:

	Rsq	KMO
radius_mean	1.00	0.81
texture_mean	0.89	0.40
perimeter_mean	1.00	0.82
area_mean	1.00	0.88
smoothness_mean	0.91	0.85
compactness_mean	0.98	0.85
concavity_mean	0.98	0.91
concave_points_mean	0.97	0.89
symmetry_mean	0.83	0.84
fractal_dimension_mean	0.95	0.88
radius_se	0.99	0.82
texture_se	0.84	0.52
perimeter_se	0.99	0.78
area_se	0.98	0.80
smoothness_se	0.82	0.62
compactness_se	0.96	0.76
concavity_se	0.97	0.74
concave_points_se	0.90	0.70
symmetry_se	0.83	0.76
fractal_dimension_se	0.94	0.78
radius_worst	1.00	0.80
texture_worst	0.94	0.39
perimeter_worst	0.99	0.85
area_worst	1.00	0.79
smoothness_worst	0.89	0.75
compactness_worst	0.97	0.75
concavity_worst	0.96	0.76
concave_points_worst	0.92	0.81
symmetry_worst	0.90	0.72
fractal_dimension_worst	0.96	0.81

In order to select only the relevant characteristics to be examined through our Principal Component Analysis, we turned to the variable-specific scores on the **Kaiser-Meyer-Olkin (KMO) Test**.
The test measures sampling adequacy for each variable in the model and is a measure of the proportion of variance among variables that might be common variance.

Following the literature, we choose to set the KMO threshold to **0.6** and we decided to exclude the following variables:

```
Pre-factor analysis diagnostics
Variable collinearity:
                    Rsq   KMO
texture_mean        0.89  0.40 *
texture_worst       0.94  0.39
texture_se          0.84  0.52
```

Moreover, the analysis of the scree plot together with the factor loadings suggested us to select **5 Principal Components** for our PCA.

* However, as the excluded variables represented the only three characteristics of the cell nuclei related to '**texture**' namely '**the standard deviation of gray-scale values**' in the digitalized images, this could potentially allow us to exclude that whole group of features from our experiment.

We could then proceed to the final analysis structured as follows:

- 5 Factors
- PCA method
- Varimax rotation
- Factor Loading > 0.6
- No double-loading

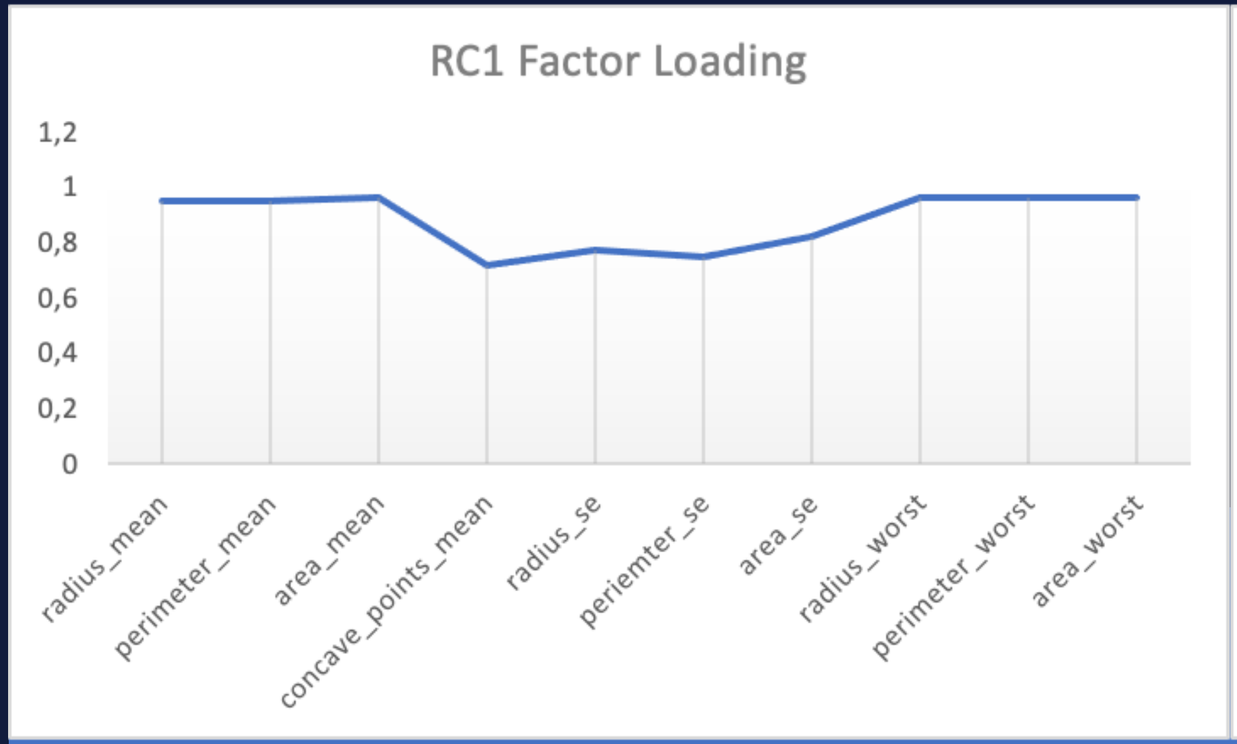
The following outputs will provide for interpretability for our results.

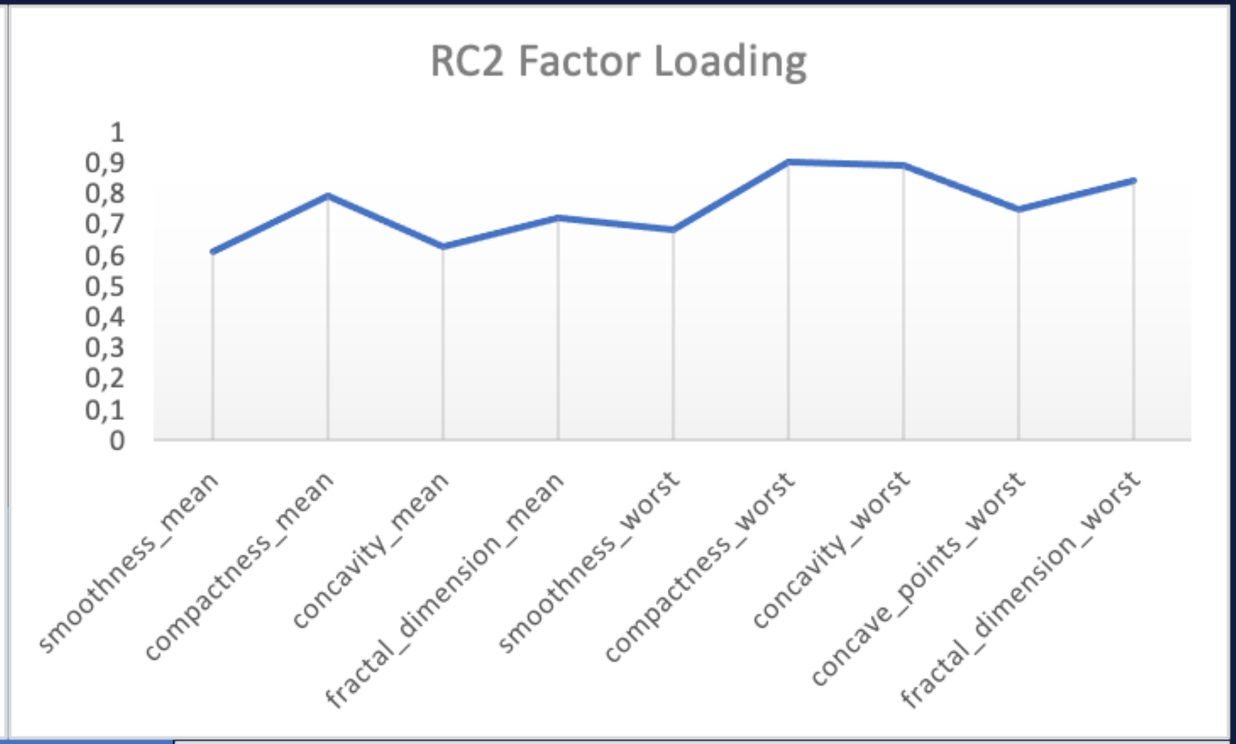
```
Factor analysis
Factors     : 5
Method      : PCA
Rotation    : varimax
Observations: 212
Correlation : Pearson

Fit measures:
                RC1  RC2  RC3  RC4  RC5
Eigenvalues    8.82 6.73 4.32 2.10 1.56
Variance %     0.33 0.25 0.16 0.08 0.06
Cumulative %   0.33 0.58 0.74 0.81 0.87
```

Our 5 factors alone explain 87% of our data, without any double-loading

Factor loadings:

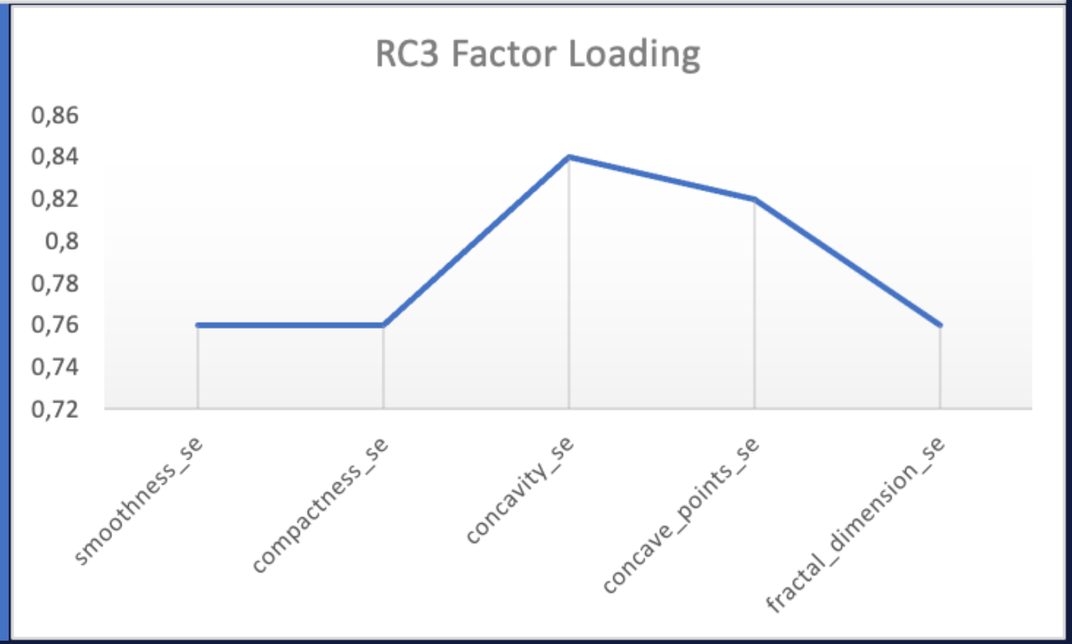
	RC1	RC2	RC3	RC4	RC5
radius_mean	0.95				
perimeter_mean	0.95				
area_mean	0.96				
smoothness_mean		0.61			0.67
compactness_mean		0.79			
concavity_mean		0.63			
concave_points_mean	0.72				
symmetry_mean				0.7	
fractal_dimension_mean		0.72			
radius_se	0.77				
perimeter_se	0.75				
area_se	0.82				
smoothness_se			0.76		
compactness_se			0.76		
concavity_se			0.84		
concave_points_se			0.82		
symmetry_se				0.79	
fractal_dimension_se			0.76		
radius_worst	0.96				
perimeter_worst	0.96				
area_worst	0.96				
smoothness_worst		0.68			
compactness_worst		0.9			
concavity_worst		0.89			
concave_points_worst		0.76			
symmetry_worst				0.73	
fractal_dimension_worst		0.84			





Factor loadings

The three displayed factors explain cumulatively 74% of our data



RC1	
radius_mean	0,95
perimeter_mean	0,95
area_mean	0,96
concave_points_mean	0,72
radius_se	0,77
periemter_se	0,75
area_se	0,82
radius_worst	0,96
perimeter_worst	0,96
area_worst	0,96

RC2	
smoothness_mean	0,61
compactness_mean	0,79
concavity_mean	0,63
fractal_dimension_mea	0,72
smoothness_worst	0,68
compactness_worst	0,9
concavity_worst	0,89
concave_points_worst	0,75
fractal_dimension_wors	0,84

RC3	
smoothness_se	0,76
compactness_se	0,76
concavity_se	0,84
concave_points_se	0,82
fractal_dimension_se	0,76

The higher the cited cell nuclei features, the higher the potential magnitude of each factor in determining **independently** the malignant nature of the breast cancer cell.

Each feature exerts an effect on the factor comparable to its loading, pictured in this case by its color gradient.

This research provided us with a useful dimensional reduction, which potentially highlighted the features explaining the highest variance in malignant breast cancer cells.

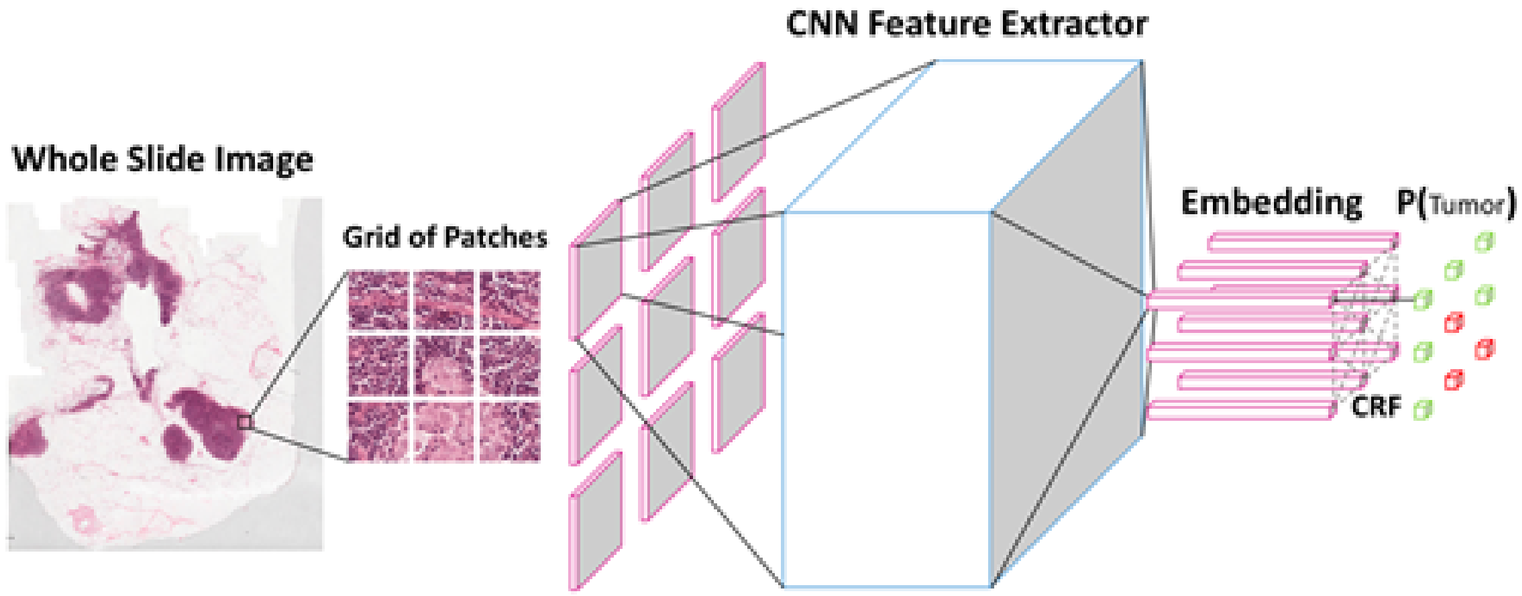
The research horizon could be improved by studying why and how each factor influences the malignant diagnosis from a clinical pathology perspective.

A glance on current horizons

Review | Open Access | Published: 27 September 2021

Deep learning in cancer diagnosis, prognosis and treatment selection

Khoa A. Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D. Williams, John V. Pearson & Nicola Waddell



Thank you!

LUDOVICO GANDOLFI
EDOARDO FERRERO