Universitat Politècnica de Catalunya

## Data Science and Engineering

# Identification of Modulators in Cancer Tumor Progression 

Bachelor's Degree Final Thesis

Advisors<br>Author<br>Dànae Canillas i Sánchez<br>Josep Vidal Manzano<br>Margarita Cabrera Bean<br>Department of Signal Theory and<br>Communications



FIB


#### Abstract

This final degree project is a study focused on the detection of relevant modulators from genetic expressions in order to help in the analysis and treatment of diagnosed breast cancer patients. The identification of the genes will be established from those that contribute in the classification of breast tumor subtypes and the ones that take part in the prediction of a possible relapse. Two methodologies will be used to find the results: Firstly, the neural networks will be used in the classification model. Then, we will apply interpretability techniques in order to provide validity to the result, allowing us to extract the pathways that have largely determined the output. Secondly, we have worked with the previous idea to compute the relapse prediction, but the results were not as good as we expected. However, two estimators that are able to model the patients relapse have been constructed. The contribution of each variable will be studied in order to establish which modulators have more prominence when modeling the risk of relapse.


Keywords: Modulators; Gene expression; Breast cancer; Neural Networks; Saliency; Prediction; Estimators

## Resum

L'objectiu d'aquest treball de fi de grau és identificar un conjunts de moduladors rellevants obtinguts a partir d'una sèrie d'expressions genètiques amb la finalitat d'ajudar en l'anàlisi i el tractament de pacients amb càncer de mama. La identificació d'aquests gens es farà analitzant un model de classificació amb l'habilitat de predir els subtipus de tumors mamaris. També formaran part d'aquest conjunt els moduladors que participen activamentent la predicció del temps de recaiguda del pacient. Els resultats s'han obtingut fent ús de dues metodologies: en primer lloc, s'utilitzaran les xarxes neuronals per dissenyar un model de classificació. A continuació, s'aplicaran tècniques d'interpretabilitat per tal de proporcionar validesa al resultats obtinguts. D'aquesta manera, es podrà identificar els moduladors que han determinat el resultat. En segon lloc, s'ha treballat amb la idea anterior per calcular la predicció del temps de recaiguda, però els resultats no han estat positius. No obstant, s'han construït dos estimadors capaços de modelar el temps de recaiguda dels pacients. Posteriorment, s'estudiarà la contribució de cada variable amb la finalitat d'establir quins moduladors han tingut més protagonisme a l'hora de modelar el risc de recaiguda.

Paraules clau: Moduladors; Expressió genètica; Càncer de mama; Xarxes neuronals; Interpretabilitat; Predicció; Estimadors

## Resumen

El objetivo de este trabajo de fin de grado es identificar conjuntos de moduladores relevantes obtenidos a partir de una serie de expresiones genéticas con la finalidad de ayudar en el análisis y el tratamiento de pacientes con cáncer de mama. La identificación de estos genes se hará analizando los resultados de un modelo de clasificación capaz de predecir los subtipos de tumores mamarios. También formarán parte de este conjunto los moduladores que participan en la predicción del tiempo de recaída del paciente. Los resultados se han obtenido a partir de dos metodologías: en primer lugar, se utilizarán las redes neuronales para diseñar un modelo de clasificación. A continuación, se aplicarán técnicas de interpretabilidad para proporcionar validez a los resultados obtenidos. De este manera, se podrán identificar los moduladores que han determinado el resultado. En segundo lugar, se ha trabajado con la idea anterior para calcular la predicción del tiempo de recaída, pero los resultados no han sido los esperados. No obstante, se han construido dos estimadores con la capacidad de modelar el tiempo de recaída de los pacientes. Posteriormente, se estudiará la contribución de cada variable para establecer qué moduladores han tenido más protagonismo en el modelado del riesgo de recaída del paciente.

Palabras clave: Moduladores; Expresión genética; Cáncer de mama; Redes neuronales; Interpretabilidad; Predicción; Estimadores

## Collaboration

This Final Degree Thesis has been developed in the context of a collaboration agreement between IDIBELL institution and the research group SPCOM (Signal Processing and Communications) from the UPC.

Biomèdica de Bellvitge


Figure 1: IDIBELL and SPCOM logotypes

Josep Vidal and Margarita Cabrera, the advisors of this project, are part of the Signal Processing and Communications Group ${ }^{1}$. It is a research team focused on the development and analysis of advanced digital signal processing techniques at the Signal Theory and Communications Department (TSC) of the Universitat Politècnica de Catalunya. Thanks to the department, we have had access to the CALCULA computing services ${ }^{2}$, needed for the training of the neural networks used in this project.

Bellvitge Biomedical Research Institute (IDIBELL) ${ }^{3}$ is a research center promoted by Bellvitge University Hospital and Viladecans Hospital focused in cancer, neuroscience, translational medicine and regenerative medicine. Miquel Àngel Pujana and Eline Blommaert are two investigators from Breast Cancer research group, whose task aims to better understand breast cancer development, the onset of subtypes and their therapeutic response and resistance. Their work is focused on deciphering and integrating the molecular mechanisms that are affected by common gene variations and rare mutations. During the evolution of the project, we have received great feedback from them: they have provided the data, helped to understand it and they also have come up with very good ideas based on their excellent experience holding up the progress of this work.

[^0]
## Acknowledgments

On the one hand, I would like to thank the dedication and interest of my advisors, Josep and Marga. It has been a pleasure to share knowledge while doing this project. I am also grateful for the confidence by IDIBELL institute; they have been always available to resolve doubts and to propose methodologies that have been beneficial in the results obtained. I was very interested in working on a project related to the medical field. I think it has been a great first experience that will lead me to be interested in similar projects in the future.

On the other hand, I want to thank my family and friends. To my parents, for teaching me to be curious and to open myself paths to freely choose what I like. To my grandparents for always being by my side and for always showing interest in what I have done, asking how an algorithm works or reading papers written in English without understanding the language. I would also like to thank the support of Maria Ribalta, I have been so lucky to meet a friend like you. And finally to Xavier Rubiés, for being there, for encouraging me and making me feel confident when facing all the challenges.

## Contents

List of Figures ..... 7
List of Tables ..... 8
1 Introduction ..... 9
1.1 Context and justification ..... 9
1.2 Motivation ..... 10
1.3 Objectives ..... 10
1.4 Work Packages ..... 11
1.5 Approach and methodology ..... 12
2 Related work ..... 13
3 Data Preprocessing ..... 14
3.1 Data Integration ..... 14
3.2 Data Exploration ..... 15
3.3 Data Cleaning ..... 18
3.4 Data Transformation ..... 20
3.5 Data Augmentation ..... 20
4 Statistical Analysis ..... 21
4.1 Hierarchical Clustering Analysis ..... 21
4.2 Data Cluster Visualization ..... 24
4.3 Correlation Study ..... 25
5 Feature Selection ..... 29
6 Experiments and results ..... 32
6.1 Tumor Class Prediction ..... 32
6.1.1 Neural Networks ..... 32
6.1.2 Interpretability ..... 37
6.2 Relapse Time Prediction ..... 40
6.2.1 Neural Network ..... 40
6.2.2 Kaplan-Meier Estimator ..... 41
6.2.3 Cox Proportional Hazard model ..... 44
6.3 User Interface ..... 49
7 Conclusions ..... 51
8 Future Work ..... 52
Glossary ..... 53
References ..... 55
Appendices ..... 58

## List of Figures

1 IDIBELL and SPCOM logotypes ..... 4
2 Gantt Chart ..... 12
3 Data preprocessing steps ..... 14
4 Variable set from the project dataset. ..... 14
5 Pathways and Immune Cells value range ..... 15
6 Age/RFS/PROLIF/Lymph Nodes Positive value range ..... 16
7 Breast tumor subtype classification ..... 18
8 KNNImputer hyperparameter $k$ tuning ..... 19
9 SMOTE technique ..... 20
10 Pathways Heatmap ..... 21
11 Two-way hierarchical clustering ..... 22
12 Pathways score values conditioned to the type of tumor. ..... 23
13 Pathway heatmap (grouped by tumor subtype) ..... 23
14 PCA and TSNE dimensionality reduction techniques ..... 24
15 Correlation heatmaps containing pathways and immune scores. ..... 25
16 Absolute difference of correlation matrices ..... 26
17 Overlap Percentage Heatmap ..... 26
18 Confidence intervals overlap ..... 27
19 Correlation heatmap of the remaining quantitative variables ..... 28
20 Normal-like class neural network training ..... 30
21 Breast cancer diagnostics components ..... 30
22 Neural Network architecture ..... 33
23 Cross-Validation process ..... 34
24 Confusion matrix validated with the test set ..... 35
25 Wrong classified patients output ..... 36
26 Saliency heatmap contribution ..... 38
27 Basal saliency sample ..... 39
28 Her2 saliency sample ..... 39
29 LumA saliency sample ..... 39
30 LumB saliency sample ..... 39
31 Kaplan-Meier Survival Curves ..... 42
32 Patient outcome based on breast tumor intrinsic subtypes ..... 43
33 Relapse modelling process ..... 45
34 Application tabs ..... 49
35 Shapley values plot from Tumor Prediction tab ..... 49
36 Immune feature scores plot from Hierarchical Clustering tab ..... 50
37 Data cluster visualization plot from Dimensionality Reduction tab ..... 50

## List of Tables

1 Distribution of the grade and stage values ..... 17
2 Missing values distribution ..... 18
3 Study of pathways correlation differences ..... 27
4 Study of pathways and immune feature scores correlation differences ..... 28
5 Use of breast tumor subtypes in different studies. ..... 29
$6 \quad$ Set of variables to be used in both models. ..... 31
$7 \quad$ Best performance networks ..... 35
8 Shapley values pattern contribution ..... 39
9 Log-Rank Test ..... 43
10 Univariate Cox regression with all the patients ..... 46
11 Univariate Cox regressions from subtype tumor based populations ..... 46
12 Basal multivariate cox regression ..... 47
13 Her2 multivariate cox regression ..... 47
14 LumA multivariate cox regression ..... 48
15 LumB multivariate cox regression ..... 48
16 Multivariate Cox regressions with all the patients. ..... 58
17 Results for cancer subtype prediction ..... 63

## 1 Introduction

### 1.1 Context and justification

The world has been paralyzed with the arrival of a global pandemic affecting the rhythm of our lives. Our attention is now focused on the evolution of the SARS-CoV-2 (COVID-19) and the administration of the vaccines, but we must not forget that there are other diseases that are also affecting population during the last year. The first wave of coronavirus infections, having occurred between the months of March and June of 2020, produced a $21 \%$ drop in cancer diagnoses in Spain, meaning that one of five cancer patients has either not been detected or has been diagnosed late. ${ }^{4}$ This will cause an increase in cancer deaths in 2021 and in the subsequent years. We cannot afford this deceleration in the breast cancer diagnosis. It is a disease that has affected thousands of women who are currently also suffering the ravages of the pandemic.

During the entire period of confinement, we have been waiting for the researchers to find a cure, but it is not an easy task. Given the fact that we live in an advanced society, all of us have come to the conclusion that it is important to do research in the health field. It is necessary to have resources to allow an automatic analysis of patient data, in order to help the doctors and the researchers who are currently collapsed at their work.

We have been fortunate to collaborate in this project with two excellent researchers: Miquel Àngel Pujana and Eline Blommaert, working in the cancer area of the IDIBELL institution. The research proposal consisted in modeling the behavior of the genes in the different subtypes of breast tumor cancer. The identification of breast cancer subgroups and their molecular drivers requires integrated views of the genome of a representative number of patients [1]. Even though classification of breast cancer is very useful in deepening breast cancer knowledge, another of the main objectives was to use the gene expression information to predict an approximation of the relapse time for each patient. As in the previous case, the researchers were interested on knowing which factors determine this risk.

[^1]
### 1.2 Motivation

We have seen how necessary it is to dedicate effort and dedication to research. We would like to contribute with new knowledge, acquired by using statistical techniques and models, that can be helpful to analyse and thus improve cancer treatment. The data science field can offer results obtained from another point of view, as many methodologies can be applied to data in order to offer results that we did not know a priori. We can find cross relationships in the variables, linear and non-linear dependencies among the data, identification of the most informative variables, etc.

> Distinguishing the signal from the noise requires both scientific knowledge and self-knowledge.
> Nate Silver

The idea is to combine the knowledge acquired from data analysis together with the experience of doctors and researchers in the sector. Usually, the health sector is forewarned with the credibility of data science results, as models seem to be mysterious black boxes. The justification of the results is important in the field of health. This is why we propose to provide explanations to all the results found, since they can be reliable and used in the treatment and detection of breast cancer. The objective is to identify the modulators that have influenced the predictions, being able to know why those variables have been decisive in the output of the models. Together with the self-knowledge of the researchers, we want to obtain conclusions sustained on the scientific basis.

### 1.3 Objectives

The milestones we set out to achieve in this project are the following ones:

- Can we identify cancer subtypes?
- Can we reliably predict the relapse time?
- Can we identify salient modulators in both cases?

The first two points focus on obtaining a classification of the different subtypes of breast cancer and a prediction of the relapse time from the genetic expressions from a set of patients. In both models, we want to provide the relevant salient features ${ }^{5}$ in each prediction, in order to justify the result.

* One of the initial objectives was to validate the results obtained with control data. This objective was modified, since we had two highly experienced researchers whose collaboration confirmed the results obtained from their knowledge and experience.

[^2]
### 1.4 Work Packages

Next, we show the division of tasks established to meet the proposed objectives. Each work package has a brief description to put the development of the proposed task in context.

1. Exploratory data analysis

Basic inspection of the dataset: range values for each variable, find possible errors/abnormal values, find missing values, compute data transformations, perform a graphical summary of the data.
2. Feature extraction

From the entire set of genes available to make a selection of those that can be used for each model.
3. Build models

- Cancer subtype prediction
- Relapse time prediction

4. Salient modulators identification

Definition of the most relevant input features applying integrated gradients.
5. Factor analysis and results evaluation

Try to find an explanation for the results obtained later evaluated by IDIBELL researchers.

The previous points correspond to the initial plan of the project. As you will see, the development of the project has had a few variations that are described below.

- Due to the obtained results of the second model, it has been necessary to expand other ways of modeling the problem. At the beginning, the idea was to work with neural networks, as it was done in the first classification problem. The obtained results were not quite good, as can it be seen in Section 6.2.1. Consequently, it was proposed to study the regression problem as a classification problem (separating the relapse time temporal axis into bins), which did not generate satisfactory results either. Finally, it was proposed to work with two different estimators. The first one (Section 6.2.2) was not parametric and it was not useful, since we wanted to study the relationships between genes and relapse risk. The other estimator was parametric and it already included the modulators that allowed us to make an analysis of the importance of each factor in the prediction (Section 6.2.3)
- Apart from the exploratory data analysis, a statistical analysis that includes correlations has been carried out, unsupervised hierarchical clustering and dimensionality reduction techniques. Some testings have been proposed by the IDIBELL team and others have been done to perceive the interaction of the variables that we were working with.


Figure 2: Gantt Chart

Figure 2 shows the project timeline represented as a Gantt chart. The work tracking was done with a weekly meeting with the advisors of the project. Additionally, we met IDIBELL researchers every two weeks to evaluate the results found.

### 1.5 Approach and methodology

In this project, we are facing two different problems. The first one, the breast tumor subtype classification, will be treated as a supervised classification task using neuronal networks. The second one will be treated as a regression problem that can be modeled in different ways. As it will be seen, neural networks, regression techniques, support vector machine and a couple of estimators have been tested in order to study the relapse risk of each patient.

All the implementations have been programmed in Python. Some of the libraries we used are: pandas for data storage and processing, sklearn to split the data and for the construction of some models, pytorch for the design and execution of neural networks, etc. However, the Cox Proportional Hazard model has been built in $R$, since the package function in this language offered better implementation features. All the results have been accompanied by graphics in order to make the output more illustrative. The libraries we used are seaborn, matplotlib and plotly.

Finally, as we will see later, an application with Dash has been developed, using graphics from the plotly library, in order to show the obtained results in an interactive and userfriendly way.

## 2 Related work

There are many publications that have focused on studying the detection of existing tumor subtypes in breast cancer. Originally, classification of breast cancer was done through the use of conventional methods such as tumor morphology, grade, and immunophenotyping for estrogen, progesterone, and HER-2 receptor expression [2]. Such techniques, that have been useful for several years, are not sufficient to accurately predict biologic behavior of breast cancer. The recently published studies want to provide a better automatic prediction to move away from the traditional pathological and clinical parameters [3] driven-decisions.

There are a wide variety of clinical trials that link genetic expressions signatures with breast cancer classification $[3,4,5]$. The technology focused on the study of these data is named gene-expression profiling. From this genetic expressions, one can observe that breast cancer is not homogeneous. One of the publications that have been taken as a reference [6] has based the classification on the activation of various oncogenic signaling genes, also named pathways. The idea is to identify tumors that share patterns of pathway activity and exhibit similar clinical and biological properties. This pattern classification was done using an unsupervised technique called two-way hierarchical clustering, widely applied in many similar studies $[7,8]$ and consequently it will be applied in this study in Section 4.1.

New technologies, such as neural networks, have provided new ways to research and diagnosis [9]. The results of these studies have started to be used by clinical oncologists in their day-to-day practice, but doctors are still determining how to use these tests and interpret them properly. The intention of this project is to use neural networks to perform the above mentioned classification with the particularity of offering the factors that have determined the result of the labeling. The idea is to justify the result by sharing the contribution of each variable in the prediction result.

There also exist other studies focused on the prognosis of each subtype [10]. Most of them subdivide the patients according to the type of tumor they have, in order to estimate the risk of failure after the surgery was done; in other words, to estimate the relapse time they may have [11, 3]. The idea is to find some prognostic and predictive factors that can predict the future occurrence of breast cancer.

Some studies assure that the best way to analyze these types of results is to use KaplanMeier estimator or the Cox Proportional Hazard model [12, 13]. These two models take into account the information provided by all the participants involved in the study, including those who dropped out during follow-up, regardless of whether they later relapsed later or not. In this project, we will work with these two estimators and we will also try to use neural networks, as in the previous problem.

## 3 Data Preprocessing



Figure 3: Preprocessing steps

This section will describe the data preprocessing techniques applied before injecting the genetic expression values into the models. In data integration (Section 3.1), we explain the source of the data being used. Next, an exploration of the data (Section 3.2) is carried out, where we describe the ranges of values for each variable in order to detect outliers. Once we have a conception of the distribution of the data, we apply some missing value detection techniques to remove the noise that can interfere in the decisions of the models (Section 3.3). Finally, some transformations will be described and computed on the dataset in order to improve the model learning (Sections 3.4, 3.5). Furthermore, the last section includes a data augmentation process to avoid data imbalance problem.

### 3.1 Data Integration

The IDIBELL researchers have provided us a processed data file from the well-known METABRIC dataset in the medical sector. The Molecular Taxonomy of Breast Cancer International Consortium database contains labeled sequencing data of 1,980 breast cancer samples. From the previous database, the institute investigators extracted a set of clinical data describing the patients. They also computed a set of pathways to be examined. The algorithm used to determine the scores for the pathways was ssGSEA (a variation on GSEA algorithm) ${ }^{6}$.


Figure 4: Variable set from the project dataset.

[^3]Additionally, they also decided to work with a set of immune feature scores. These variables were extracted from the expression data of the METABRIC dataset by using the CIBERSORTx algorithm ${ }^{7}$, a tool that extracts the immune cell related content in each tumor sample. Furthermore, additional variables have been included since IDIBELL researchers think they can help us to improve the results of the model. All of them can be found in the Glossary. We include a summary in the Figure 4.

### 3.2 Data Exploration

In the database, there are a total of 1986 patients described by a set of 53 variables, 14 of them are pathways scores and 22 are immune feature scores, computed by the previous mentioned algorithms. In this section, we will start by observing the distribution of values of the pathways and the immune feature scores:

## Continuous Features



Figure 5: The first figure represents the range of values from each pathway. The second one shows the ranges of the immune cell data. The graph used is a range plot, since we want to see where most of the data is concentrated in order to detect the outliers in Section 3.3.

[^4]As it can be seen on the upper chart in Figure 5, the values of the pathways are normalized between -0.5 and 0.5 , and the mean of all the variables is very similar. Otherwise, the values of the variables corresponding to the immune feature scores only take positive values between 0 and 0.5 . According to IDIBELL researchers, the pathways and variables corresponding to the immune feature scores do not present any outlier, since the ssGSEA and the CIBERSORTx output returns values normalized between 0 and 1 .

In Figure 6, we will see the range of values corresponding to the other continuous variables. The distribution of each one has been conditioned to the breast tumor subtype. In this dataset, we have five tumor subtypes: LumA, LumB, Her2, Basal-like and Normal-like. The Normal-like subtype is not included in the plots because it will not be taken into account in the models (see the explanation in Section 5).


Figure 6: Representation of the ranges of values from Age, RFS, PROLIF, Lymph Nodes Positive variables conditioned by the breast tumor subtypes.

Looking at the first plot in Figure 6, we can see that the age at diagnosis distribution for each type of tumor is very similar, the values are within the life expectancy ranges that are currently established in society. We can find very young patients, as we find one around 20 years old. We can also appreciate elderly patients (note the point corresponding to a patient having an age close to 100 years).

The relapse time plot (RFS variable boxplot) shows the relapse time for patients with $R F S E=1$, i.e. the relapsed patients. We do not take into account the patients with $R F S=0$ because it only informs of the duration of the study, since the patient has not relapsed. We see that the time averages are quite dispersed among the different tumor subtypes. The patients classified as Basal-like seem to have the earliest relapse time.

As for the PROLIF variable, we see that there are clearly two different distributions, the $\operatorname{LumA}$ subtype has smaller values compared to the distribution of the other tumor types. As IBIDELL team stated, the values are normalized between -1 and 1 , therefore we do not find any value outside the range established as normal. Finally, in the number of positive lymph nodes plot we can see some extreme values that appear far away from the average of the variable. These strange values have been reviewed by the IDIBELL researchers and they have assured us that they were not outliers, since the real range of number of positive lymph nodes variable already includes the previous values.

## Categorical Features

Next, we will the range of values for categorical variables:

- Integrative clusters (INTCLUST): $\{1,2,3,5,6,7,8,9,10,4 E R-, 4 E R+\}$
- Cellullarity: $\{$ Low, Moderate, High $\}$
- Treatment: It is a variable that takes combinations of these three strings: $\{H T, R T, C T\}$, corresponding to the treatment administered.
- TP53.mut and PIK3CA.mut: $\{W T, M U T\}$
- Nottingham Prognostic Index (NPI): is the product of three categorical factors (see the formula in the Glossary). It takes values between 1 and 7 .

The integrative clusters variable is a second classification that METABRIC provides apart from the classification by subtypes. Then, we have the Cellullarity variable that indicates the amount of cancer cells in the tumor. We also find two variables (TP53.mut and PIK3CA.mut) that take binary values depending on whether the pathways they describe presents mutations or not. Last, the Treatment variable indicates the type of treatment the patient is receiving based on the type of tumor it has. Finally, we examine the grade and stage variables:
Grade:

|  | 1 | 2 | 3 |
| :--- | :--- | :--- | :--- |
| LumB | $3,81 \%$ | $41,10 \%$ | $55,08 \%$ |
| Her2 | $1,74 \%$ | $25,32 \%$ | $72,92 \%$ |
| LumA | $17,61 \%$ | $56,62 \%$ | $25,76 \%$ |
| Basal | $0,09 \%$ | $9,40 \%$ | $89,65 \%$ |


| Stage: |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| 0 | 1 | 2 | 3 | 4 |
| $27,68 \%$ | $32,44 \%$ | $33,68 \%$ | $4,96 \%$ | $1,24 \%$ |
| $24,15 \%$ | $34,74 \%$ | $32,20 \%$ | $8,89 \%$ | - |
| $35,83 \%$ | $32,86 \%$ | $28,19 \%$ | $2,83 \%$ | $0,28 \%$ |
| $19,87 \%$ | $31,15 \%$ | $40,06 \%$ | $8,56 \%$ | - |

Table 1: Distribution of the grade and stage values conditioned to tumor subtypes.

On the one hand, the variable grade can take three different values. We see that the value 3 is the most common in the majority of tumor subtypes, except for Her2, where the majority is established in patients whose grade value is 2 . The least represented value in all subtypes is grade $=1$. On the other hand, regarding the stage variable, we see that the values are quite balanced in the values 0,1 and 2 . The percentage decreases when the value of the variable increases stage $=3$. Finally, we can see that stage takes a value equal to 4 only with the luminal types.


Figure 7: Breast tumor subtype classification
To conclude with this section we show a pie chart in Figure 7 representing the distribution of tumor subtypes. We see that the predominant class is $L u m A$ and the least represented is Her2 (without taking into account the Normal-like subtype).

### 3.3 Data Cleaning

This section is focused on the treatment of missing values in order to work with a consistent database.

## Missing values

Table 2 shows the distribution of missings values. We can see that the variable with the highest number of missings is stage. The pairs of variables RFS/DSS10 and RFSE/DSSE10 have the same number of missings. This is because they are the same metric scaled differently, therefore the missings are found in the same patients.

| Variable | \# missings |
| :--- | :--- |
| RFS | 3 |
| DSS10 | 3 |
| RFSE | 20 |
| DSSE10 | 20 |
| grade | 88 |
| stage | 448 |
| lymph_nodes_positive | 7 |
| NPI | 1 |
| CELLULARITY | 68 |
| INTCLUST | 5 |

Table 2: Distribution of the missing values. On the left, the variables having missings and on the right the number of patients that have missing value in the corresponding variable.

Next, we will describe the decisions taken to deal with the missing values:

- Due to the nature of the Integrative clusters variable and the fact that it contains a small number of missing values (a total of 5 patients, who also have missing in other variables) we decided to drop this five samples.
- The patients who had three or more variables with missing values were also removed.

Finally, with the remaining patients, the KNNImputer ${ }^{8}$ technique has been used in order to fill the missing fields. The KNNImputer method offered by the scikit-learn library is one of the most used techniques to impute missing values.

The idea behind this procedure is to use the $k$-Nearest Neighbors ( $k N N$ ) algorithm to find neighboring observations in order to estimate the missing values from them. The process consists in computing the mean value of the samples to estimate the value of the missing cell. This algorithm requires the number of neighbors $(k)$ to be used to impute the unknown value. It is a hyperparameter that must be optimized. To do this, a routine has been created that tests a RandomForestRegressor using cross-validation with different values of $k$. In Figure 8 we show that the optimal value of neighbors has been $k=7$.


Figure 8: KNNImputer hyperparameter $k$ tuning. For each value of $k$ we represent the root mean square error (RMSE) of the trained model

[^5]
### 3.4 Data Transformation

There is a set of variables needed to be transformed in order to work properly with them later. The first transformation was a normalization of the dataset values in order to set the ranges of all the variables between 0 and 1 . The remaining modifications are focused on encoding the categorical string variables to categorical integer variables:

- Treatment: The treatment variable has been divided into three binary variables: $H T, R T$ and $C T$ corresponding to the three types of treatments that a patient can receive. A 1 has been assigned to each variable if the corresponding therapy was in the treatment variable.
- Cellularity: $\left\{{ }^{\prime}\right.$ Low' $^{\prime}$, 'Moderate ${ }^{\prime},{ }^{\prime}$ High' $\}$ have been recoded to $\{1 / 3,2 / 3,1\}$.
- TP53.mut and PIK3CA.mut: This variable manifests the possibility of a gene mutation, for TP53 and PIK3CA separately, as $\left\{{ }^{\prime} W T^{\prime}=0,{ }^{\prime} M U T^{\prime}=1\right\}$


### 3.5 Data Augmentation

As it can be seen in Figure 7, we have a data class balancing problem, since we do not have a balanced distribution among classes. In fact, we have three times more data from patients categorized as $\operatorname{LumA}$ than Her2 diagnosed patients. Due to the difference in proportions in the labeled data of each subtype we used the SMOTE technique:

## SMOTE

One approach to addressing imbalanced datasets is to oversample the minority classes. The most widely used approach to synthesizing new examples from tabular data is called the Synthetic Minority Oversampling TEchnique, or SMOTE for short [14].

The process of generating new synthetic data is simple: for each patient samples of the minority class, we find its $k$ nearest neighbors (by default $k=5$ ). Then, using any distance metric, the difference between the feature vector from the minority class sample and its neighbors is calculated. Finally, this difference is multiplied by any random value between $(0,1]$ and is added to the previous feature vector, obtaining a synthetic sample (Figure 9).


Figure 9: SMOTE technique operation example

## 4 Statistical Analysis

In this section, some statistical analysis have been developed to anticipate what we may find in the results of the models. The idea is to explore the existence of patterns in the data, to detect clusters, to find redundant variables, etc.

### 4.1 Hierarchical Clustering Analysis

The intrinsic subtypes Luminal A (LumA), Luminal B (LumB), Her2 and Basal-like have been extensively studied by microarray and hierarchical clustering analysis. [15, 16, 17, 18]. In this project, it has been decided to also carry out this unsupervised analysis in order to obtain a classification based on the values of each variable, like the first publications did. For this study, only the pathways variables have been taken into account.


Figure 10: Pathways Heatmap

In Figure 10, the values for each pathway are represented by a color scale. The values are between -0.5 and 0.5 . In warmer tones, we find the positive values and in colder tones, the negative values. It is a very illustrative graph, since it allows to see the value distribution for each individual. The horizontal axis has been ordered according to the tumor subtypes of the patients. From left to right, we find Basal first, followed by Her2, LumA, LumB and finally Normal-like breast tumor subtypes. On the vertical axis, we have the set of pathways with which the study will be carried out.

We could draw some conclusions: the most informative pathways in this classification seem to be the ERBB2 and ESSR1. For the Basal type, we find very low ESR1 values. For the Her2 type, we have very high values in ERBB2. On the other hand, in $\operatorname{LumA}$ and $L u m B$, the high values are found in ESR1. Finally, in Normal-like tumors, it seems that there is no factor that differentiates it from other tumors.

## Two-way hierarchical clustering

The two-way hierarchical clustering is a technique widely used in health data visualization to see the intrinsic classification that each patient has, based on the values of the factors studied. The idea is to build a bottom-up ${ }^{9}$ hierarchy of clusters on both axes: at the beginning, each observation starts in its own cluster. Then, pairs of clusters are merged one by one moving up in the dendogram ${ }^{10}$ lines. The splits are determined by the linkage method: the data is grouped according to the correlation of each groups, as we can see in the formula:

$$
\begin{equation*}
\text { correlation coefficient }=1-\frac{(u-\bar{u}) \cdot(v-\bar{v})}{\|(u-\bar{u})\|_{2} \cdot\|(v-\bar{v})\|_{2}} \tag{1}
\end{equation*}
$$

where $\bar{v}$ is the mean of the elements of group vector $v$, and $x \cdot y$ is the dot product of $x$ and $y$. The dendrogram on the horizontal axis classifies patients with similar pathway values. On the vertical axis, we carry out this classification according to the behavior of the pathways.


Figure 11: Two-way hierarchical clustering

To improve the understanding of the result, we colored the first row of the heatmap by the tumor subtype classification (Basal: 'violet', Her2: 'black', Normal: 'pink', LumA: 'blue', LumB: 'green'). In Figure 11, we can observe a clear differentiation between two categories: one that contains patients classified as Basal and a second category that contains the

[^6]Luminal patients (both $A$ and $B$ ). We can also see that the Her2 and Normal-like subtypes are scattered within the other breast cancer tumor groups.

A boxplot is attached below (Figure 12) to see the distribution of values for each tumor subtype. We can see, as it has been verified in the previous plot, that there exist a value dispersion of the ESR1 and ERBB2 pathways.

Pathways boxplot conditioned by PAM50


Figure 12: Pathways score values conditioned to the type of tumor.

To make the differences between pathways more understandable, we decided to group the patients by breast tumor subtype computing the mean value of each class, as can be seen in Figure 13:


Figure 13: Pathway heatmap (grouped by tumor subtype)

As we concluded in the previous graphs, we see that the behavior of the pathways of patients diagnosed with Luminal breast cancer subtype are very similar, as they have high values in ESR1. Otherwise, the patients categorized as Basal-like have small values in the previous variable, but they have quite high averaged values in the Cell Cycle variable. Regarding the Her2 tumor, we observe that it takes high values in a variable little affected in the previous classes, the ERBB2 modulator. Finally, regarding Normal-like patients, we did not find extreme values in any of the pathways.

In Figure 13, we also added a column corresponding to the unlabeled patients of the database. These samples can be used in the future to obtain the predictions in both models. At first glance, it seems that on average they have values similar to the breast tumor Luminal subtype.

### 4.2 Data Cluster Visualization

Twenty years ago, a study led by Perou et al [15] determined the existence of various subtypes of cancer, specifying 6 types: Luminal A, B, and C, Her2, Basal-like and Normallike. The results were extracted from an unsupervised analysis of gene expression profiles over a 65 breast cancer surgical tissues. In this work, we have tried to carry out the same process with the available genes (review them in Section 3).


Figure 14: Sample of the two dimensionality reduction techniques, on the left Principal Component Analysis (PCA) and on the right t -distributed Stochastic Neighbor Embedding (t-SNE)

We are facing a problem with multiple variables and we want to understand what relationship exists among them. To do this, the first step has been the reduction of the number of variables to two dimensions. In the case of PCA [19], these two dimensions correspond to the two most informative components, with the highest variance. Otherwise, the t-SNE [20] estimates probability distributions of the data.

In Figure 14, we see two large clouds of points and we do not identify any separate set from the total. Each plot includes two data distribution diagrams on both axes and the dots have been colored according to the tumor subtype. On the one hand, we clearly see that the Basal class is quite separated from the others.

On the other hand, the luminals are concentrated in the same space, without differentiating the class $A$ and $B$. Patients diagnosed with Her2 are more separated in the t-SNE plot from the rest. Finally, we see that the Normal-like group appears very dispersed. As in Figure 13, we decided to add unlabeled patients to see if they could be categorized into any of the classes. By the dispersion of the points we could categorize the majority of them in the luminal groups, as we stated in Section 4.1.

### 4.3 Correlation Study

During the development of the project, the IDIBELL researchers were interested in studying two different groups of patients: those who relapsed early and those who did not relapse after several years. In this research, we compute the correlations between the pairs of variables from each population. Next, we will see two heatmaps with the cross correlation of each population for each pair of variables:


Figure 15: Correlation heatmaps containing pathways and immune scores. On the left, the relapsed patients in two years. In the left figure, not relapsed patients in five years

In the left heatmap in Figure 15, we observe higher absolute values (dark-colored cells) than in the right one. At first glance, it is very difficult to conclude whether these differences are significant. For this reason, the absolute difference between the two correlation matrices has been calculated.


Figure 16: Absolute difference of correlation matrices

In Figure 16, the differences between the heatmaps can be observed more clearly. The darker cells represent the variable correlation pairs with the greatest difference between the two populations studied. To verify that these differences are statistically significant, we have computed the respective confidence intervals. To affirm if two pairs have a significant difference, we define two hypotheses:

- Null Hypothesis: The null hypothesis states that there is no significant difference between the two correlation pairs calculated in each population
- Alternative Hypothesis: There is a clear difference between the two correlation values obtained in each population.

If there is no overlap between the two confidence intervals, the correlation difference is considered significant. Otherwise, we cannot affirm the previous case.

In Figure 17, we show the percentage of overlap between the confidence intervals obtained in the previous step. To verify which pairs of variables have a significant correlation difference, we have to search if there is a match between the darkest points in Figure 16 and those in Figure 17.


Figure 17: Overlap Percentage Heatmap

To facilitate the comparison, we attach two tables: Table 3 shows the ten pairs of pathways with the greatest correlation difference. In Table 4, we show the same results obtained in the latter table, but now we are including the immune feature scores.

| Var1 | Var2 | diff | corr_R2 | corr_NR5 | CI $(90 \%)$ Var1 | CI $(90 \%)$ Var2 | \% Overlap |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| MYC | WNT | 0.3663 | 0.3807 | 0.0144 | $[0.2079,0.5305]$ | $[-0.0376,0.0663]$ | 0.0 |
| HIPPO | ERBB2 | 0.3368 | -0.285 | 0.0517 | $[-0.4487,-0.1029]$ | $[-0.0002,0.1034]$ | 0.0 |
| WNT | ERBB2 | 0.3115 | -0.3991 | -0.0877 | $[-0.5459,-0.2286]$ | $[-0.139,-0.0359]$ | 0.0 |
| RTK_RAS | SRC | 0.2725 | 0.29 | 0.0174 | $[0.1082,0.453]$ | $[-0.0345,0.0693]$ | 0.0 |
| SRC | ERBB2 | 0.2429 | -0.1892 | 0.0537 | $[-0.364,-0.0016]$ | $[0.0017,0.1053]$ | 0.0 |
| HIPPO | ESR1 | 0.238 | -0.5471 | -0.3091 | $[-0.6664,-0.4006]$ | $[-0.3554,-0.2614]$ | 0.0 |
| TGF.Beta | Hypoxia | 0.2266 | -0.1532 | 0.0733 | $[-0.3314,0.0355]$ | $[0.0215,0.1248]$ | 3.07 |
| NOTCH | ESR1 | 0.2167 | -0.1646 | 0.0521 | $[-0.3417,0.0239]$ | $[0.0002,0.1038]$ | 5.32 |
| TGF.Beta | SRC | 0.2114 | 0.4318 | 0.2204 | $[0.2656,0.573]$ | $[0.1704,0.2692]$ | 0.89 |
| MYC | SRC | 0.2066 | 0.2978 | 0.0912 | $[0.1166,0.4598]$ | $[0.0394,0.1424]$ | 6.15 |

Table 3: Study of pathways correlation differences.

The first two columns in Table 3 correspond to the pairs of variables to be compared. The diff column has the absolute correlation difference between both populations: those who relapse in less than two years (corr_R2) and those who do not relapse in five years (corr_NR5). Below, we show four different plots (Figure 18) with the percentage of overlap between the confidence intervals of some pairs of pathways appearing in Table 3.


Figure 18: Confidence intervals overlap examples from Table 3.

|  |  | diff | corr_R2 | corr_NR5 | CI $(90 \%)$ Var1 | CI (90\%) Var2 | \% Overlap |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| NK cells activated | Macrophages M0 | 0.444 | -0.0651 | 0.3789 | $[-0.2497,0.1241]$ | $[0.3335,0.4225]$ | 0.0 |
| NOTCH | T cells CD4 memory act. | 0.4212 | -0.3983 | 0.0228 | $[-0.5453,-0.2277]$ | $[-0.0292,0.0747]$ | 0.0 |
| Macrophages M1 | Mast cells activated | 0.4056 | 0.1665 | -0.2392 | $[-0.0219,0.3434]$ | $[-0.2875,-0.1896]$ | 0.0 |
| NK cells resting | Macrophages M0 | 0.3708 | 0.172 | -0.1988 | $[-0.0162,0.3485]$ | $[-0.2482,-0.1484]$ | 0.0 |
| MYC | WNT | 0.3663 | 0.3807 | 0.0144 | $[0.2079,0.5305]$ | $[-0.0376,0.0663]$ | 0.0 |
| Hypoxia | Plasma cells | 0.3601 | 0.2932 | -0.067 | $[0.1116,0.4558]$ | $[-0.1185,-0.0151]$ | 0.0 |
| HIPPO | ERBB2 | 0.3368 | -0.285 | 0.0517 | $[-0.4487,-0.1029]$ | $[-0.0002,0.1034]$ | 0.0 |
| ERBB2 | Monocytes | 0.3238 | 0.3265 | 0.0027 | $[0.1478,0.4845]$ | $[-0.0493,0.0546]$ | 0.0 |
| WNT | ERBB2 | 0.3115 | -0.3991 | -0.0877 | $[-0.5459,-0.2286]$ | $[-0.139,-0.0359]$ | 0.0 |
| Dendritic cells activated | Neutrophils | 0.3107 | 0.2894 | -0.0213 | $[0.1076,0.4525]$ | $[-0.0732,0.0307]$ | 0.0 |

Table 4: Study of pathways and immune feature scores correlation differences.

In Table 4, all the correlation differences are significant because any of the confidence intervals overlap. These results have been delivered to the IDIBELL team and they will study the significant variable pairs behaviour in each population.

## Auxiliary variables



Figure 19: Correlation heatmap of the remaining quantitative variables

Finally, the correlations between the remaining quantitative variables have been calculated. In Figure 19, we can see that there is a high correlation between the RFS and DSS10 variables, as well as the RFSE and DSSE10 variables. The relationship between the latter variables will be discussed in Section 5 , where we will see they contain the same information. There are two other variables that show a high correlation: the NPI and grade. This is because the first variable is calculated as follows:

$$
\begin{equation*}
\boldsymbol{N P I}=0.2 \cdot[\text { Size of the index lesion }]+[\text { Node status }]+[\text { Grade of tumour }] \tag{2}
\end{equation*}
$$

## 5 Feature Selection

After analysing the available variables in the datasets, it is necessary to do a pruning in the dataset to only work with features that can be useful.

## Normal-like subtype

As it can be seen in the exploratory and statistical analysis, it seems that the Normal-like subclass does not show any pattern in the modulator values. This fact was discussed with the IDIBELL researchers and they stated that the samples of this subclass were not quite clear, due to the large presence of normal tissue in the tumors.

There are other studies that have also decided not to use the last mentioned subtype [21, $18]$. One of the studies found in the literature [15] mentions there is no certainty whether the normal breast-like group truly represents breast cancer tissue or was merely a sampling error of the benign breast tissue embedded in breast cancer tissue.

In fact, the TCGA dataset, one of the most used in this type of research, also discards this subtype tumor class, as can be seen in the Table 5:

| Study | $\begin{aligned} & \text { Data } \\ & \text { type } \end{aligned}$ | Breast tum or subtype |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Luminal |  | HER2 <br> over-expression | TNP |  | Normal-like |
|  |  | Luminal A | Luminal B |  | Basal | Non-basal |  |
| Sørlie [12] | GEP | 1 | 1 | 1 | 1 |  | 1 |
| Hu [17] | GEP | 1 | 1 | 1 | 1 |  | 1 |
| PAM50 [18] | GEP | 1 | 1 | 1 | 1 |  | 1 |
| Abd 日-Rehim [23] | TMA | 1 | 1 | 1 | 1 |  | 1 |
| Sotiriou [24] | GEP |  |  | 1 | 2 |  |  |
| Fan [25] | GEP | 1 | 1 | 1 | 1 |  |  |
| Lehmann [26] | GEP |  |  |  | 2 | 4 |  |
| Gatza [87] | PATH | 2 | 5 | 2 | 3 |  |  |
|  |  | 4 |  |  |  |  |  |
|  |  | 1 |  |  |  |  |  |
| METABRIC [88] | INT | 3 | 3 | 1 | 1 |  |  |
|  |  |  |  |  |  |  |  |
|  |  | 1 |  |  |  |  |  |
| TOGA [89] | INT | 1 | 1 | 1 |  |  |  |
| Dai [33] | INT | 1 | 1 | 1 |  |  |  |

Table 5: This table represents the use of breast tumor subtypes in each dataset. The gray blocks indicate if the tumor subtype class is available. [22]

The remaining four intrinsic subtypes of breast cancer, luminal $A$ and $B$, Her2 and Basallike, are widely accepted in most studies. In addition, it was decided to check if the neural network was able to capture any pattern of the data that can discriminate the Normal-like class from the others, but as it can be seen in Figure 20, the results were not satisfactory. We can see that the percentage of accuracy of the last mentioned class is too small, both in the training confusion matrix and in the validation one.


Figure 20: Normal-like class neural network training. On the left, the confusion matrix corresponding to training. On the right, the same confusion matrix with the validation data.

As we previously mentioned, the training of the neural networks will be carried out in the next sections. In Figure 20, we only show the result of one of the trainings executed before discarding this class.

## Correlated variables

In the statistical analysis, we found that some sets of variables have a lot of correlation. Correlated variables can introduce redundant information to the models. In Figure 19, we have seen that there exist a high correlation between DSS10, DSSE10 and RFSE and $R F S$. In fact, they contain the same information but the first pairs are scaled to 10 years. Consequently, the first two variables will not be considered in the following models.

## Other cases

Next, we will explain why some of the dataset variables have been discarded:

- The INTCLUST variable defines a molecular classification different from the one we are studying, as it is shown in Figure 21. The mentioned feature will not be considered in both studies, since it does not conform to the proposed objectives.


Figure 21: Breast cancer diagnostics components [21]

- The Treatment variable is conditioned to the tumor breast cancer subtype. Therefore, this variable will not be included in the training.


## Selected features

In Table 6, we show the set of variables that will be considered in the models.

| Immune feature scores | Pathways | Other features |
| :--- | :--- | :--- |
| B cells naive | Cell Cycle | PAM50 |
| B cells memory | HIPPO | RFS |
| Plasma cells | MYC | RFSE |
| T cells CD8 | NOTCH | PROLIF |
| T cells CD4 naive | NRF2 | Grade |
| T cells CD4 memory resting | PI3K | Stage |
| T cells CD4 memory activated | TGF.Beta | Number of lymph nodes positive |
| T cells follicular helper | RTK_RAS | Treatment |
| T cells regulatory | TP53 | TP53.mut |
| T cells gamma delta | WNT | PIK3CA.mut |
| NK cells resting | Hypoxia | Nottingham Prognostic Index |
| NK cells activated | SRC | Cellularity |
| Monocytes | ESR1 |  |
| Macrophages M0 | ERBB2 |  |
| Macrophages M1 |  |  |
| Macrophages M2 |  |  |
| Dendritic cells resting |  |  |
| Dendritic cells activated |  |  |
| Mast cells resting |  |  |
| Mast cells activated |  |  |
| Eosinophils |  |  |
| Neutrophils |  |  |

Table 6: Set of variables to be used in both models. The variables used as target are highlighted in bold.

## 6 Experiments and results

In this section, we will detail the models built in the project. The first model is a classification neural network. The classification task will consist in distinguishing between the different subtypes of tumors using the features selected in Section 5. The second model is focused on estimating the relapse time. Since we have a limited set of relapsing patients, we have tested different methods. We will begin by describing some trials carried out with neural networks and then we will explain two estimators that have been able to model the stated problem.

### 6.1 Tumor Class Prediction

The first objective of this project was to find out if it is possible to predict tumor subtypes with the available data. We are facing a supervised classification problem that takes the variable PAM50 as a target. This variable has four possible values: Basal, Her2, LumA and $\operatorname{LumB}$, the tumor subtypes in which we can classify the patients in the database. It has been decided to use neuronal networks because we proposed to carry out a saliency study to provide interpretability to this type of models. (see Section 6.1.2).

### 6.1.1 Neural Networks

In the design of the neural network architecture, we only have contemplated to use fullyconnected layers (FCNNs). The main characteristic of this structure is that all the nodes (or neurons) in one layer are connected to the neurons in the next layer. One of the advantages of using this architecture is that it does not impose any structure. In other words, it does not make assumptions about the data or task the network will perform. In this kind of network, we need to set some hyperparameters, that are described below:

## Architecture

1. The first step was to select the number of hidden layers that the neural network should have. We have tested networks with one, two, three and four hidden layers. The best results were obtained with a network containing three hidden layers.
2. Once the number of layers had been chosen, we have tested different models varying the number of neurons for each layer (in the previous step, the tests were made with different numbers of neurons as well). We tried to increase the number of neurons with respect to the input, increasing the number of dimensions of the feature space and also reducing the number of neurons in each layer gradually until reaching four neurons in the output layer. The network that works best had the following configuration (see Figure 22):


Figure 22: Neural Network architecture

We can see that the number of neurons in each layer gradually decreases. The first layer corresponds to the number of input variables (the second feature selection will be detailed in the training step). The hidden layers have 30,20 and 10 neurons respectively. Finally the last layer corresponding to the ouput contains 4 neurons, the number of classes of the target variable.

In the output of the last layer an activation function has been added. The two most used activation functions in deep learning are the sigmoid and the ReLU. During the training we have used both, but he ReLu is the non-linear function that has worked best.

## Training

The training of the neural network has been done by using the cross-validation technique. Cross-validation is a resampling procedure that has helped us to evaluate the network training using different data partitions for validation. In all the tests, we have set the number of folds to 5 .

The cross-validation trials allowed us to define the set of variables that would enter to the final model. We tested all the possible combinations of variables for each model and we have chosen the set of variables with higher validation accuracy averaged over all folds. The process is described in Figure 23.


Figure 23: Cross-Validation process

The selected set of variables in the final model contains the 14 pathways, the 22 immune feature scores, the PROLIF variable and the stage and grade features.

## Regularization

It has been decided to introduce some regularization parameters to avoid the memorization of the network and thus to be able to generalize the learning. In pytorch there are several optimizers. We decided to use the Adam optimizer ${ }^{11}$ since it is the one that offers the best performance. This optimizer, based on the stochastic gradient descent technique, offer the tuning of the following parameters:

- Learning rate ( $\alpha$ parameter): indicates the weight update proportion. Larger values results in faster initial learning. Smaller values slow learning during the training. We have tested $\{1 e-2,1 e-3\}$ values.
- Epsilon: it is used to prevent any division by zero in the implementation. The set of tested values has been $\{1 e-8,1 e-6\}$.
- L2-penalty: we add a penalty equal to the square of the coefficients magnitude. L2 will not yield sparse models and none of the coefficients will be eliminated (main difference with L1-penalty). We tested this parameter with $\{1 e-4,1 e-6\}$ values.
- Dropout: consists of randomly set to 0 some proportion of the neurons that feed into a fully connected layer during each step of gradient descent. It helps to reduce model overfitting.


## Results

In Table 17 (see Apendices) you can find the hyperparameters tuning and different trials varying the number of epochs and batch sizes. The best results are shown in Table 7, including the accuracy for the test set:

[^7]| $\#$ | lr | epochs | batch_size | dropout | eps | weight_decay | $\operatorname{Tr}$ | Val | Te |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 78 | 0.001 | 80 | 200 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 87.0745 | 82.9787 | $\mathbf{8 3 . 5 8 9 7}$ |
| 79 | 0.001 | 100 | 100 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 89.0957 | 84.2553 | 83.4188 |

Table 7: Best performance networks

The best models had a learning rate equal to 0.001 . The majority achieved a validation accuracy greater than $80 \%$. It seems that the network did not need much L2-penalty: the models which have worked better have been those with the lowest value of weight decay (0.0001). The introduction of the dropout has been positive because it has helped to avoid overfitting.
In paper [15], the prediction results stated that the easiest class to predict was $L u m A$ patients, followed by LumB and Her2 patients, and Basal-like patients did the worst.


Figure 24: Confusion matrix validated with the test set.

Thanks to data augmentation techniques, we have obtained a fairly balanced accuracy for all classes. The worst class to predict has been Her2, followed by LumB. Surprisingly, the Basal-like class has been the class with the least prediction errors.

We also generated a plot (see Figure 25) containing the wrong classified patients. In fact, we show the same information that has been provided in the confusion matrix of Figure 24. However, in this plot, we are adding the classification percentage for each instance. Each column represents a patient and it is divided into four rows, each one representing a tumor subtype where we colored the percentage of belonging to each class. The percentages are obtained by computing the softmax function to the output of the network.


Figure 25: Wrong classified patients output

We can see that the majority of Basal patients have been incorrectly assigned to Her2. On the other hand, the Her2 are wrongly classified as LumA class. Finally, we see that the network has problems differentiating the luminal patients.

Once the network has been trained with the learning set (see Figure 23) and having tested it with the test set (Table 7), we proceed to apply the saliency techniques in order to explain the modulators contribution in the predictions.

### 6.1.2 Interpretability

One of the main objectives of this project was to try to understand how much each feature of the neural network contributed to the prediction. The neural networks have been considered veritable black-boxes some years ago. However, in recent years, most of the research has focused on providing interpretability to machine learning and deep learning models in order to understand why certain decisions have been made. As Doshi-Velez and Kim stated [23], the accuracy of a model is a metric that offers an incomplete description of most real-world tasks. Regarding to this project, focused on predicting medical factors, the need to justify the results obtained is a must.

## Model-Agnostic Methods

The widely used specific models like linear regression, decision trees, etc are interpretable at a modular level since they can be classified according to whether they are linear, monotonous, their type of output, etc. However, there exist another type of models, named agnostic models. The agnostic methods are independent of all the last mentioned characteristics, they are flexible to adapt their results to any type of model. Such models are responsible for separating the explanations from the machine learning model. One of the best known methods, SHapley Additive exPlanation model [24], will be used to understand the prediction obtained with the neural network.

## $\underline{\text { Shapley Values }}$

As it was said at the beginning of the section, we want to know the contribution that each variable has in the prediction. The shapley values, extracted from SHapley Additive exPlanation method, indicate the distribution of knowledge contribution that each factor has. It is computed as the average marginal contribution of a variable across all the possible coalitions of the available features. The shapley values interpretation is the following:

The value of the $j$-th feature contributed $\phi_{j}$ to the prediction of this particular instance compared to the average prediction for the dataset. [25]
and it is calculated as

$$
\begin{equation*}
\phi_{j}(v a l)=\sum_{S \subseteq\left\{b_{1}, \ldots, b_{n}\right\} \backslash\left\{b_{j}\right\}} \frac{|S|!\cdot(n-1-|S|)!}{n!}\left(\operatorname{val}\left(S \cup\left\{b_{j}\right\}\right)-\operatorname{val}(S)\right) \tag{3}
\end{equation*}
$$

where $\phi_{j}$ is the Shapley value for $j$-th feature $\left(b_{j}\right)$ for one sample $x$ in the dataset. $S$ is the subset of the features used in the model. We sum all the coalitions in $S$ that do not contain the feature $b_{j}$.

The $\left[f_{j}\left(b_{S} \cup b_{i}\right)-f_{j}\left(b_{S}\right)\right]$ factor is the marginal contribution that the feature $b_{j}$ has when it is added to $S$ coalition. This factor is multiplied by $|S|!\cdot(n-1-|S|)$ !, the number of permutations of $S$. Then, the result is divided by $n!$ in order to average the result over all the possible orderings of $x$.


Figure 26: Saliency heatmap contribution

In Figure 26, we show a heatmap with the corresponding shapley values for each patient. The horizontal axis has been ordered according to the tumor subtypes. The first row indicates the predictions made by the neural network. The green colored patients correspond to the samples whose prediction coincides with the target. Otherwise, the red colored ones are associated to wrong classified patients (to facilitate comparison, the second row has been colored according to the tumor subtypes of each patient). The following rows correspond to the shapley values contribution to each variable (conditioned to the average prediction for the dataset). In the last rows of the figure, we can see the darkest variables, those which represent the highlighted shapley values in the prediction.

We observe that the immune feature scores have had little contribution to determine the output of the network. In fact, they do not show large absolute values to be considered relevant in the prediction. Looking at the pathways features, we see that ESR1 and ERBB2 are the determinant features in the patients classification. On the other hand, the variables PROLIF, stage and grade have also provided a lot of information in the prediction.

| Shapley Values Contribution |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
|  | ESR1 | ERBB2 | PROLIF | Stage | Grade |
| Basal | High Positive | - | High Positive | Medium Positive | Low Positive |
| Her2 | Medium Positive | Medium Negative | High Positive | Medium Positive | Low Positive |
| LumA | High Negative | - | - | Low Positive | - |
| LumB | High Negative | - | High Positive | Medium Positive | - |

Table 8: Shapley values pattern contribution

In Table 8, we included the features having the largest contributions to the network output. We described the average level of contribution they had in each tumor subtype.


Figure 27: Basal saliency sample

Figure 29: LumA saliency sample


Figure 28: Her2 saliency sample

Figure 30: LumB saliency sample

In Figures 27, 28, 29 and 30, four patients have been randomly sampled, each one from a different tumor subtype. We can appreciate the contribution of each variable looking at the heights of the bars of each plot. For example, we observe that the Basal patient has the variables ESR1, PROLIF and stage with a positive contribution. If we go back to Figure 26, we can observe the same Basal pattern: the highlighted shapley values have a positive contribution in the last mentioned variables. We see a similar pattern in Her2 patients but including the negative contribution of $E R B B 2$ feature. Regarding to luminal patients, we see that the set of variables that contribute to the prediction are the practically the same.

### 6.2 Relapse Time Prediction

In this second section, we will try to model the relapse time for each tumor subtype. To do this, the RFSE and RFS variables will be used. The RFS variable quantifies the time that elapses from the subject's entry into the study until a certain event occurs. In this study the event is the relapse of the patient, reported by the RFSE variable. We started working with different neural networks arquitectures (Section 6.2.1). Nevertheless, we will end up working with two estimators widely used in the medical sector (Sections 6.2.2 and 6.2.3).

### 6.2.1 Neural Network

As in the previous classification problem, we have tried to model the relapse time by training a neural network. The idea was to formulate the problem by using two models. The first one attends to classify a patient according to the possibility of relapse. If the network predicts the relapse of a patient, this sample will be injected to a second model to predict relapse time. The problem had to be separated into two models because it does not make sense to train the time that a patient stays in the study (does not relapse) as the relapse time. The main drawback was that the number of relapsed patients was approximately 450 and it is not efficient to train a network with so many variables and so little instances.

- The training of the first model was quite good. In the best case, we have obtained a training accuracy of $74.36 \%$, but a test accuracy of $62 \%$. The class with the best prediction was the relapsed patients population, with $79.67 \%$ of accuracy in the test dataset. The network structure that worked best had three hidden layers, as the one used in subtype classification, using the same set of variables.
Considering the problem as a binary classification, different methodologies were also tested such as support vector machines (with and without applying kernels) and decision trees. The support vector machine was not able to find a plane separating the two classes without making many mistakes, and with decision trees we obtained an accuracy of the $58.46 \%$.
- Regarding the prediction of the relapse time, we trained a network with the entire set of relapsed patients. Several architectures and hyperparameters were tested, but the network was unable to learn. As the results of the network were not good, it was decided to try with simpler models: linear regression, logistic regression, regression using kernels introducing non-linear combinations of the data, etc. The results were not good either, and this led us to look for other alternatives of modeling.


### 6.2.2 Kaplan-Meier Estimator

Since the neural networks did not work as expected, we decided to explore other methods in order to obtain better results that could relate the pathways with the prediction of the time relapse. We started with a widely used model in the health analytic field introduced by Edward L. Kaplan and Paul Meier in 1958 [26].

The main objective of the Kaplan-Meier method is to build a survival function estimator $S(t)$. It is a non-parametric method, since it does not assume any probability function (as we do not know its real distribution). The estimator is obtained by maximizing the likelihood function of the population sample. It takes into account the elements described below:

- An $N$ independent and identically distributed sample of the population, that defines the number of patients in the study.
- Each patient has a time measure ( $R F S$ variable), with a label ( $R F S E$ ) based on whether or not they have relapsed.
- The patients at risk are the elements of the sample waiting to an event. In this particular case, they do not know if they will relapse or not.
- The time axis will be divided into $k$ partitions $(k \leq N) t_{1}<t_{2}<\ldots<t_{k}$.
- At every partition $t_{i}$, there are $n_{i}$ patients at risk and $d_{i}$ relapse events are observed.
- The discrete hazard rate $h_{i}$ is the probability of an individual to relapse in time $t_{i}$

In this problem we want to find the maximum likelihood estimation of $h_{i}$, we will name it $\hat{h}_{i}$. We define the survival rate ${ }^{12}$ as :

$$
\begin{equation*}
S(t)=\prod_{i: t_{i} \leq t}\left(1-h_{i}\right) \tag{4}
\end{equation*}
$$

The likelihood is then

$$
\begin{equation*}
\mathcal{L}\left(h_{j: j \leq i} \mid d_{j: j \leq i}, n_{j: j \leq i}\right)=\prod_{j=1}^{k} h_{j}^{d_{j}}\left(1-h_{j}\right)^{n_{j}-d_{j}} \tag{5}
\end{equation*}
$$

To build this function, it has been assumed that for each relapse, the event occurs after the relapse is observed. Therefore the log-likelihood is

$$
\log (\mathcal{L})=\sum_{j=1}^{k}\left(d_{j} \log \left(h_{j}\right)+\left(n_{j}-d_{j}\right) \log \left(1-h_{j}\right)\right)
$$

[^8]To find the maximum we need to derivate with respect to $h_{i}$

$$
\begin{aligned}
\frac{\partial \log (\mathcal{L})}{\partial h_{i}} & =\frac{d_{i}}{\hat{h}_{i}}-\frac{n_{i}-d_{i}}{1-\hat{h}_{i}}=0 \\
& \Rightarrow \hat{h}_{i}=\frac{d_{i}}{n_{i}}
\end{aligned}
$$

and replacing in the survival function we obtain

$$
\begin{equation*}
\hat{S}(t)=\prod_{i: t_{i} \leq t}\left(1-\frac{d_{i}}{n_{i}}\right) \tag{6}
\end{equation*}
$$

We have decided to build a Kaplan-Meier fitter for each subtype of tumor. The time intervals have not been sampled uniformly. Each time interval only includes a relapsed event. In other words, parameter $d_{i}$ will always be equal to 1 because only one patient relapses in each $t_{i}$.


Figure 31: Kaplan-Meier Survival Plot. The plot above shows the four survival curves corresponding to the population of each tumor subtype. The plot below includes the number of patients at risk of relapse in each time period. Since there are few samples of relapsed patients in the database, the IDIBELL researchers recommended to focus only in the first five years of study of each population.

Looking at Figure 31, it seems that the curve of LumA is further away from the others. To verify if this difference is significant, we will use the widely used Log-Rank Test [27], which formulates the following two hypotheses:

- Null Hypothesis: There is no difference between the survival curves
- Alternative Hypothesis: The survival curves are different:

| Log-Rank Test | Log rank (Chi-Square) | df | p-value |  |
| :--- | :--- | :--- | :--- | :--- |
| LumA | LumB | 37.46 | 1 | $\mathbf{9 . 3 2 e - 1 0}$ |
| LumA | Her2 | 56.02 | 1 | $\mathbf{7 . 1 6 e - 1 4}$ |
| LumA | Basal | 43.88 | 1 | $\mathbf{3 . 4 8 e - 1 1}$ |
| LumB | Her2 | 4.64 | 1 | $\mathbf{0 . 0 3 1 1}$ |
| LumB | Basal | 1.51 | 1 | 0.2198 |
| Her2 | Basal | 0.93 | 1 | 0.3359 |

Table 9: Log-Rank Test. In bold, p-values $<0.05$ are highlighted

Computing the test, we can see in Table 9 that the p-value is less than 0.05 in all the LumA combinations, meaning there is a significant difference between the groups compared. We also observe that the populations corresponding to Her2 and Basal are significantly different from each other.

As discussed in some publications $[15,10]$, the patients diagnosed with $\operatorname{LumA}$ have the best overall relapse free survival. $L u m B$ and Her2 patients have intermediate outcomes, and Basal-like patients are the ones having a higher risk of relapse in less time. In Figure 32 , there is a graphic extracted from a publication [22] that relates the patient outcome with breast tumor intrinsic subtypes. The authors describe the same curves behavior that we have in Figure 31.


Figure 32: Patient outcome based on breast tumor intrinsic subtypes [22]

### 6.2.3 Cox Proportional Hazard model

Once we had the results of the previous model, the IDIBELL researchers proposed the Cox Proportional Hazard model since it seemed interesting to apply the contribution of the pathways in a model that could include covariates. It is a regression widely used by medical researchers to investigate the association between the survival time of patients and one or more predictor variables set instead of a single variable at a time, like in the Kaplan-Meier model. As in all regressive models, we will obtain a set of coefficients for each pathway and thus we will be able to see what contribution each variable has in the prediction. We define the hazard function as

$$
\begin{equation*}
H(t)=h_{0}(t) \cdot \exp \left(\beta_{1} x_{1}+\beta_{2} x_{2}+\ldots \beta_{n} x_{n}\right) \tag{7}
\end{equation*}
$$

where $t$ is the survival time, $x_{i}$ are the covariates and $\beta_{i}$ are the coefficients that measure the impact of covariates. The values $\exp \left(\beta_{i}\right)$ are called the hazard ratios (HR). If the hazard ratio in a variable is greater than 1 indicates that as the value of it covariate increases, the event hazard increases, and thus the duration of survival decreases. However, if it is less than 1 indicates that the event hazard decreases, therefore the duration of survival increases. To work with this model, some modifications on the data have been needed and are described below:

## Data Factorization

The IDIBELL researchers recommended us to split the continuous variables into some bins in order to obtain more representative results (internally, the R-function computes too many partitions of the continuous variables and each bin contained a small set of unrepresentative samples). Next, we will detail the bins generation by each variable:

- Pathways and : are variables whose values are normalized between -0.5 and 0.5 . The values have been discretized as follows:

$$
\text { Genes variables } \begin{cases}\text { Low } & \text { if } x \in[-0.5, \text { mean }(v a r)],  \tag{8}\\ \text { High } & \text { if } x \in(\text { mean }(\text { var }), 0.5]\end{cases}
$$

- age at diagnosis: it is a variable that takes values between 20 and 100 (in the METABRIC dataset). Since the range is wide, it has been decided to make three partitions of the data:

$$
\text { Age at diagnosis } \begin{cases}\text { Young } & \text { if } x \in(20,45]  \tag{9}\\ \text { Adult } & \text { if } x \in(45,70] \\ \text { Elderly } & \text { if } x \in(70,100]\end{cases}
$$

- number of positive lymph nodes: usually the prognostic significance is evaluated by counting the number of positive lymph nodes. In publication [28] the authors define
a high risk from 10 nodes, the same threshold has been established in our data:

$$
\text { Lymph positive nodes } \begin{cases}\text { Low } & \text { if } x \in(0,10],  \tag{10}\\ \text { High } & \text { if } x \in(10,+\infty)]\end{cases}
$$

- The stage and grade variables were already discrete. The first variable takes values between 0 and 4 , otherwise the grade variable range is between 1 and 3 .

The meaning of the $\beta$ coefficients will be conditioned by the reference bin for each variable. To make the comparison easier and following the recommendations from IDIBELL researchers, the reference bin has been established as the bin that contains the smallest values (Low bins, Young bin defined by age at diagnosis variable, 0 bin for stage, etc).

## Modeling

The procedure for the construction of the models is divided into several steps:

1. First, we will start by constructing some univariate models, that estimate the relapse time as a function of each single variable in the dataset.
2. These models will be estimated over different sets of populations. The first population will consist in the entire set of patients available in the database. The other populations will correspond to the patients classified according to their tumor subtype. In other words, a model will be built for the whole population corresponding to the Basal, Her2, luminal $A$ and luminal $B$ subtype tumor patients.
3. For each population, the set of variables that have been significant in the modeling of the relapse time will be extracted. Afterwards, we will model a multivariate cox regression for each set of features. In this second modeling, all possible interactions between the variables will be included. Then, we can identify the prognostic factors according to the sign of the $\beta$ coefficients of the significant variables.


Figure 33: Relapse modelling process

## Univariate Cox Regression

Each row in Table 10 represents an univariate model using the variable indicated in the Variable column. Only those variables that have been significant in their corresponding model appear in the table (the features which have had a p-value equal to or less than 0.05).

We see that the variables can take negative or positive coefficients. If the $\beta$ coefficient is positive indicates that as the value of the variable increases, the event probability increases, and thus the duration of survival decreases. However, if it is less than 0 indicates that the event hazard decreases, therefore the duration of survival increases.

The Table 10 has been computed dividing the total set of samples into four populations corresponding to the tumor subtypes.

| Variable | $\beta$ | p-value |
| :--- | :--- | :--- |
| T.cells.CD4.naive | -0.29537 | 0.0018 |
| Cell_Cycle | 0.34238 | 0.000283 |
| HIPPO | 0.24169 | 0.00984 |
| NRF2 | -0.38125 | $5.52 \mathrm{e}-05$ |
| Hypoxia | 0.20804 | 0.0265 |
| SRC | 0.25439 | 0.00715 |
| ESR1 | -0.57299 | $8.15 \mathrm{e}-10$ |
| ERBB2 | 0.2489 | 0.00764 |
| age_at_diagnosis (Adult) | -0.4765 | 0.000219 |
| age_at_diagnosis (Elderly) | -0.3140 | 0.031446 |
| PROLIF | 0.8048 | $2.29 \mathrm{e}-15$ |
| lymph_nodes_positive | 1.4389 | $<2 \mathrm{e}-16$ |
| NPI | 1.16938 | $<2 \mathrm{e}-16$ |
| stage (1) | 0.6861 | 0.000318 |
| stage (3) | 1.02325 | $1.52 \mathrm{e}-09$ |

Table 10: Univariate Cox regressions with all the patients.

| Variables | Basal |  | Her2 |  | LumA |  | LumB |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\beta$ | p-value | $\beta$ | p-value | $\beta$ | p-value | $\beta$ | p-value |
| Macrophages.M0 | -0.4027 | 0.0486 |  |  |  |  |  |  |
| lymph_nodes_positive | 1.1148 | 0.000305 | 1.4556 | $2.32 \mathrm{e}-06$ | 2.1969 | $2.65 \mathrm{e}-07$ |  |  |
| NPI | 0.7766 | 0.000104 | 0.7509 | 0.000359 | 0.6876 | 0.000491 | 0.9627 | $6.67 \mathrm{e}-09$ |
| stage (3) | 0.8717 | 0.00663 | 0.8822 | 0.0135 | 0.8373 | 0.0120 | 0.7109 | 0.0343 |
| Dendritic.cells.activated |  |  | 0.4829 | 0.0191 |  |  |  |  |
| PROLIF |  |  |  |  | 0.4955 | 0.0109 | 0.3804 | 0.0216 |
| ERBB2 |  |  |  |  | 0.4654 | 0.0183 |  |  |
| SRC |  |  |  |  | 0.6668 | 0.00136 | 0.3719 | 0.0261 |
| Eosinophils |  |  |  |  | \| -0.4699 | 0.0199 |  |  |
| T.cells.regulatory.Tregs. |  |  |  |  | \| 0.3921 | 0.0449 |  |  |
| NRF2 |  |  |  |  |  |  | -0.4364 | 0.00865 |

Table 11: Univariate Cox regressions from four populations, corresponding to tumor subtypes patients classification. The first column corresponds to the set of variables that have been significant in each population. The following columns are the population studied, indicating the $\beta$ coefficients and the p-value obtained.

We observe that in all the models the NPI and stage variables have been relevant in modeling the time of relapse. Another variable that has also taken on great importance is the number of positive lymph nodes. By comparing Table 10 with Table 11, we observe that the five studied groups (all the patients and each subtype tumor populations) share the last three mentioned variables as relevant in the relapse estimation.

## Multivariate Cox regression analysis

The multivariate models will include the set of relevant variables interactions. For example, if we have three variables $x 1, x 2, x 3$ the regression formulation will be:

$$
\begin{aligned}
\mathrm{y} & =\mathrm{b} 0+\mathrm{b} 1 * \mathrm{x} 1+\mathrm{b} 2 * \mathrm{x} 2+\mathrm{b} 3 * \mathrm{x} 3+\mathrm{b} 4 *(\mathrm{x} 1 * \mathrm{x} 2)+\mathrm{b} 5 *(\mathrm{x} 1 * \mathrm{x} 3)+\mathrm{b} 6 *(\mathrm{x} 2 * \mathrm{x} 3) \\
& +\mathrm{b} 7 *(\mathrm{x} 1 * \mathrm{x} 2 * \mathrm{x} 3)
\end{aligned}
$$

We obtain a coefficient for each variable and we also include the interaction between all combinations of features. Since the number of combinations can be high, we will restrict the number of variables to be considered in each model to five. The variables pruning in the models that had a greater number of features has been set by IDIBELL. Additionally, since these models work with the interaction of variables, we will consider the continuous version of the variables in order to compute them.

We start with the population group that includes all patients. Table 16, in the Appendices, corresponds to the estimation of the relapse time using a multivariate model for the entire patient dataset. We can see that there are too many significant factors in the results table. This is because we work with continuous variables. However, we are not so interested in the coefficients since they are not consistent with this modification, so now we will observe the sets of relevant variables. All the combinations that contain the variable number of positive lymph nodes have been outstanding. Since this feature appeared in the previous models we can lead to the conclusion that it is an important variable in the prediction of relapse time. Next, we will see which sets of variables are relevant in the populations corresponding to the classification by tumor subtypes:

| Variable | beta | p-value |
| :--- | :--- | :--- |
| lymph_nodes_positive | 1.32521 | 0.1285 |
| NPI | 0.72632 | $\mathbf{0 . 0 1 3 1}$ |
| stage | 0.99434 | 0.2153 |
| lymph_nodes_positive*NPI | -0.22448 | 0.1180 |
| lymph_nodes_positive*stage | -1.21281 | 0.0352 |
| NPI*stage | -0.21889 | 0.1858 |
| lymph_nodes_positive*NPI*stage | 0.20726 | $\mathbf{0 . 0 2 7 2}$ |

Table 12: Basal multivariate cox regression

| Variable | beta | p-value |
| :--- | :--- | :--- |
| lymph_nodes_positive | 1.01702 | 0.143 |
| NPI | 0.28617 | 0.230 |
| stage | 0.04708 | 0.955 |
| lymph_nodes_positive*NPI | -0.14794 | 0.202 |
| lymph_nodes_positive*stage | -1.21281 | 0.305 |
| NPI*stage | -0.36013 | 0.833 |
| lymph_nodes_positive*NPI*stage | 0.05312 | 0.366 |

Table 13: Her2 multivariate cox regression

We used the same set of variables in multivariate Basal (Table 12) and Her2 (Table 13) models. We observe that, in the Basal model NPI factor and the interaction of NPI, stage and number of positive lymph nodes variables are significant features, whereas the Her2 population does not highlight any variable.

| Variables | beta | p-value | Variables | beta | p-value |
| :---: | :---: | :---: | :---: | :---: | :---: |
| lymph_nodes_positive | 0.98947 | 0.0599 | lymph_nodes_positive | 1.16351 | 0.1366 |
| NPI | 0.68228 | 0.0118 | NPI | 0.95179 | 0.0161 |
| stage | 1.07702 | 0.2798 | stage | 0.65492 | 0.5456 |
| PROLIF | -2.94811 | 0.1990 | PROLIF | 8.51151 | 0.0191 |
| lymph_nodes_positive*NPI | -0.20633 | 0.0465 | lymph_nodes_positive*NPI | -0.23324 | 0.1328 |
| lymph_nodes_positive*stage | -0.37760 | 0.1583 | lymph_nodes_positive*stage | -0.41278 | 0.3051 |
| NPI*stage | -0.37050 | 0.1083 | NPI*stage | -0.13664 | 0.5831 |
| lymph_nodes_positive*PROLIF | 2.27668 | 0.1029 | lymph_nodes_positive*PROLIF | -0.85707 | 0.6283 |
| NPI*PROLIF | 0.92392 | 0.1489 | NPI*PROLIF | -2.04314 | 0.0251 |
| stage*PROLIF | 3.00777 | 0.1815 | stage*PROLIF | -3.65626 | 0.1685 |
| lymph_nodes_positive*NPI*stage | 0.09775 | 0.0483 | lymph_nodes_positive*NPI*stage | 0.08645 | 0.2848 |
| lymph_nodes_positive*NPI*PROLIF | -0.50050 | 0.0628 | lymph_nodes_positive*NPI*PROLIF | 0.33210 | 0.3505 |
| lymph_nodes_positive*stage*PROLIF | -1.29949 | 0.0773 | lymph_nodes_positive*stage*PROLIF | 0.54922 | 0.5680 |
| NPI*stage*PROLIF | -0.56377 | 0.3248 | NPI*stage*PROLIF | 0.89124 | 0.1411 |
| lymph_nodes_positive*NPI*stage*PROLIF | 0.27338 | 0.0549 | lymph_nodes_positive*NPI*stage*PROLIF | -0.18080 | 0.3493 |

Table 14: LumA multivariate cox regression. Table 15: LumB multivariate cox regression.

The Tables 14 and 15, corresponding to the luminal patients, have taken also the same sets of features. We can obsesrve that the NPI variable appears to be significant another time as part of the interaction with the other variables and individually.

## Results analysis

NPI is one of the most relevant variables in relapse time prediction [29]. It has come as a relevant variable in all the studied models, as well as the number of positive lymph nodes variable. Other variables that are used in the literature as factors in modeling are: patient age, tumor size, tumor stage and number of involved auxillary lymph nodes [3]. In fact, the age at diagnosis factor and the stage features have been relevant in the dataset we are working with. All the cancer studies that use the beta coefficients in modeling the time to relapse or the death event establish the following relationship:

- A covariate with $\beta>0$ is called bad prognostic factor
- A covariate with hazard ratio $\beta<0$ is called good prognostic factor

The features that appeared to be relevant in most of the univariate models have positive betas, they are factors that define a bad prognosis.

### 6.3 User Interface

In order to make the results more accessible, an interactive web application has been created. Most of the plots that appear in this work have been built by the Python Plotly library. Using the Dash package from the same authors, a space has been developed to interact with the visualizations presented.

## Identification of modulators in cancer tumor progression

Welcome $\quad$ Hierarchical Clustering $\quad$ Dimensionality Reduction $\quad$ Correlation Study $\quad$ Tumor Prediction $\quad$ Relapse Probability Prediction

Figure 34: Application tabs

The application has 6 window tabs (see Figure 34):

- Welcome tab: contains a description of the project.
- Hierarchical Clustering tab: includes the study carried out in Section 4.1.
- Dimensionality Reduction: includes the study done in Section 4.2.
- Correlation Study tab: shows the plots that appear in Section 4.3.
- Tumor Prediction tab: includes the results of the patient classification model described in Section 6.1.
- Relapse Probability Prediction tab: includes the results of the two estimators used in Section 6.2.

Next, we show some screenshots of the application:


Figure 35: Shapley values plot from Tumor Prediction tab. The plot shows the shapley values from a randomly chosen patient. The user is able to select the patient tumor subtype.

## Immune cells



Figure 36: Immune feature scores plot from Hierarchical Clustering tab. The user can select the features that appear in the heatmap, in order to facilitate the pattern comparison.


Figure 37: Data cluster visualization plot from Dimensionality Reduction tab. The user can filter the data shown in the above two plots.

The interface is publicly available in:

- Identification of modulators in cancer tumor progression GitHub Repository


## 7 Conclusions

Gene expression profiling has given us insight into the molecular complexity of breast tumors. In this work, we found that, from a relatively small set of genes, it has been possible to build a model able to classify the patients based on their tumor subtype. This classification was already obtained by doctors and researchers. From the genes that they considered relevant, they were capable of diagnosing the tumor subtype for each patient. The main advantage of using the neural network model proposed in this project is that we are able to know which the contributions of each variable in the tumor subtype classification are. The diagnoses are usually offered by checking whether the values of a certain set of variables are within empirically established ranges. Thanks to the shapley values, we are able to know which variables have determined the output of each patient, offering their corresponding contribution. In this way, we can detect the influence of factors that have also appeared significant in the two-way hiererchical clustering heatmaps. For example, we have seen that the ESR1 and ERBB2 pathways are clearly decisive in the classification of the tumor type; or in the opposite case, it seems that the immune feature scores do not play a relevant role in the prediction.

The other objective of the project was to identify a set of features that was relevant in predicting the relapse time of patients. It is known, that in most applications, neural networks are the models that obtain the best results, since they are able to find non-linear relationships between data that other techniques cannot capture. In this study, we have worked with a small set of samples of relapsed patients. In the data exploration, we have seen that the relapse time was different for the different types of tumor. Having such heterogeneous dataset of relapsed patients has not facilitated the use of neural networks. It was necessary to look for other alternatives to identify which variables had more weight when predicting the time of relapse.

One of the best known estimators for modeling the relapse time of cancer patients is the Kaplan-Meier model. The Kaplan-Meier plots have been very useful to reaffirm that the relapse time is conditioned by the tumor subtype of the patients, since the curves for each population were quite different. We can conclude that the patients diagnosed with $\operatorname{LumA}$ are the least likely to relapse, whereas Her2 and Basal patients have the worst prognosis. The downside of modeling the problem with the previous model is that it cannot be parameterized with the information extracted from the genetic expressions. To solve the problem working with covariates, the IDIBELL team proposed to use the Cox proportional hazard models. With these models, we estimated the contributions of the introduced covariates. The univariate models have been tested to perform the feature selection used in the multivariate models. From these models, we observed that NPI, stage and the number of positive lymph nodes were very influential when predicting the relapse time.

The results obtained in this project have been reaffirmed with the researchers knowledge. The challenge of using the proposed models is to provide helpful tools in order to have different points of view when it comes to diagnosing patients and improving their treatment.

## 8 Future Work

The time invested to the realization of this project has allowed us to obtain the whole set of results described in the previous sections. Even so, in the near future, we would like to work on the following ideas:

- Try to obtain more data from relapsed patients. In this project, only a group of around 450 patients were available. When we have few samples and many variables that describe them, it is difficult to train a complex model such as a neural network.
- One of IDIBELL's proposals was to calculate the coefficients of the cox proportional model with a neural network in order to include nonlinear feature interactions. If we had more data it would be possible to do better training to get an approximation of the coefficients, as discussed in the previous point.
- Apply the study of variables to other types of cancer. One of the most used datasets is the Cancer Genome Atlas. TCGA is a coordinated effort from the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) decided to give an understanding of the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing. The dataset contains more than 11,000 tumour samples analysed from 33 cancer types. Doing an analysis of this dataset could be very interesting, both for the insights that can be discovered about each cancer and also to be able to make comparisons between the different genetic expressions.
- With the results obtained in the correlation study (see Section 4.3), the IBIDELL team has studied the pairs of variables having the most differences between the patient population relapsed in less than two years and the set of patients that did not relapse before five years. It would be very enriching to participate in the progress of this study.
- App improvement: provide more functionalities to the application, such as adding model training options, user-defined display parameters (for example, coloring the dimensionality reduction plot by the target decided by the user). It is also considered adding a visualization of the internal functioning of the neural network (evolution of loss, state of gradients, activation of neurons, etc).


## Glossary

DSS10 : disease specific survival (in days) at 10 years
Values: Temporary continuous variable
Pages: . 18, 28, 30
DSSE10 : disease specific survival event at 10 years
Values: $\{0,1\}$
Pages: . 18, 28, 30
INTCLUST : breast cancer classification with distinctive molecular profiles and clinical outcomes, complementary classification to PAM50 variable.
Values: $\{1,2,3,5,6,7,8,9,10,4 E R-, 4 E R+\}$
Pages: . 17, 30
METABRIC : the Molecular Taxonomy of Breast Cancer International Consortium database is a Canada-UK Project which contains targeted sequencing data of 1,980 breast cancer samples.
Pages: . 14, 17
NPI : Nottingham Prognostic Index $=[0.2 \mathrm{x}$ size of the index lesion $]+$ node status + grade of tumour
Pages: . 17, 28, 47, 48, 51
PAM50 : it is a 50-gene signature that classifies breast cancer tumor subtypes.
Values: $\{$ LumA, LumB, Her2, Basal-like, Normal-like $\}$
Pages: . 32
PIK3CA.mut : indicates if the sample contains a mutation in PI3K pathway.
Values: $\left\{{ }^{\prime} W T^{\prime},{ }^{\prime} M U T^{\prime}\right\}$
Pages: . 20
PROLIF : cell proliferation is how quickly a cancer cell copies its DNA and divides into 2 cells. If the cancer cells are dividing more rapidly, it means the cancer is faster growing or more aggressive.
Values: $\in[0,1]$
Pages: . 17, 38, 39
$\boldsymbol{R F S E}$ : indicates the occurrence of relapse
Values: $\{0,1\}$
Pages: . 17, 18, 28, 30, 41
$\boldsymbol{R F S}$ : indicates the time from diagnosis to occurrence of event (in days)
Values: Temporary continuous variable
Pages: . 17, 18, 28, 30, 41
TP53.mut : indicates if the sample contains a mutation in TP53 pathway.
Values: $\left\{{ }^{\prime} W T^{\prime},{ }^{\prime} M U T^{\prime}\right\}$
Pages: . 20
age at diagnosis : age when the patient was diagnosed with breast cancer

Values: $\in[20,100]$
Pages: . 16, 44, 45, 48
grade : based on how much the cancer cells look like normal cells. ${ }^{13}$
Values: $\{1,2,3\}$
Pages: . 17, 18, 28, 34, 38, 45
immune feature scores : scorings associated with immunity genes
Values: $\in[0,0.5]$
Variables: 'B cells naive', 'B cells memory', 'Plasma cells', 'T cells CD8', 'T cells CD4 naive', 'T cells CD4 memory resting', 'T cells CD4 memory activated', 'T cells follicular helper', 'T cells regulatory (Tregs)', 'T cells gamma delta', 'NK cells resting', 'NK cells activated', 'Monocytes', 'Macrophages M0', 'Macrophages M1', 'Macrophages M2', 'Dendritic cells resting', 'Dendritic cells activated', 'Mast cells resting', 'Mast cells activated', 'Eosinophils', 'Neutrophils' Pages:. 15, 16, 34, 38
number of positive lymph nodes : guides the treatment and helps predict prognosis, the more lymph nodes that contain cancer, the poorer prognosis tends to be.
Values: Continuous variable
Pages: . 17, 44, 47, 48, 51
pathways : the molecular pathways describe a series of actions among molecules in a cell that leads to a certain end point or cell function.
Values: $\in[-0.5,0.5]$
Variables: 'Cell_Cycle', 'HIPPO', 'MYC', 'NOTCH', 'NRF2', 'PI3K', 'TGF.Beta', 'RTK_RAS', 'TP53', 'WNT', 'Hypoxia', 'SRC', 'ESR1', 'ERBB2'
Pages: . 13-17, 21, 23, 24, 27, 34, 38, 41, 44, 51
stage : classifies the breast cancer based on how much cancer there is in the body and where it is when first diagnosed. ${ }^{14}$
Values: $\{0,1,2,3,4\}$
Pages: . 17, 18, 34, 38, 39, 45, 47, 48, 51
treatment : treatment received by the patient, the variable takes combinations of the three possible values.
Values: $\left\{{ }^{\prime} R T^{\prime}\right.$ (radiotherapy $),{ }^{\prime} C T^{\prime}($ chemotherapy $),{ }^{\prime} H T^{\prime}$ (hormonal) $\}$
Pages: . 20

[^9]
## References

[1] Jonnathan G Santillán Benítez, Á Quiroz-Ordóñez, Hugo Mendieta Zerón, and Leobardo Gómez-Oliván. Expresión génica y receptores hormonales en cáncer mamario. "el camino hacia la búsqueda de terapias preventiva". Revista de Medicina e investigación, 1:17-24, 012013.
[2] Bhuvaneswari Ramaswamy Robert Wesolowski. Gene expression profiling: Changing face of breast cancer classification and management. Gene Expression The Journal of Liver Research, 15:105-115, 2011.
[3] Grazia Arpino, Daniele Generali, Anna Sapino, Lucia Del Matro, Antonio Frassoldati, Michelino de Laurentis, Paolo Pronzato, Giorgio Mustacchi, Marina Cazzaniga, Sabino De Placido, Pierfranco Conte, Mariarosa Cappelletti, Vanessa Zanoni, Andrea Antonelli, Mario Martinotti, Fabio Puglisi, Alfredo Berruti, Alberto Bottini, and Luigi Dogliotti. Gene expression profiling in breast cancer: A clinical perspective. The Breast, 22(2):109-120, 2013.
[4] Maryann Kwa, Andreas Makris, and Francisco Esteva. Clinical utility of geneexpression signatures in early stage breast cancer. Nature reviews. Clinical oncology, 14, 052017.
[5] Christos Sotiriou, Soek-Ying Neo, Lisa Mcshane, Edward Korn, Philip Long, Amir Jazaeri, Philippe Martiat, Stephen Fox, Adrian Harris, and Edison Liu. Breast cancer classification and prognosis based on gene expression profiles from a populationbased study. Proceedings of the National Academy of Sciences of the United States of America, 100:10393-8, 092003.
[6] Michael Gatza, Joseph Lucas, William Barry, Jongwook Kim, Quanli Wang, Matthew Crawford, Michael Datto, Michael Kelley, Bernard Mathey-Prevot, Anil Potti, and Joseph Nevins. A pathway-based classification of human breast cancer. Proceedings of the National Academy of Sciences of the United States of America, 107:6994-9, 03 2010.
[7] Andrea Bild, Guang Yao, Jeffrey Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Andrew Berchuck, John Olson, Jeffrey Marks, Holly Dressman, Mike West, and Joseph Nevins. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature, 439:353-7, 022006.
[8] Linda Vidman, David Källberg, and Patrik Rydén. Cluster analysis on high dimensional rna-seq data with applications to cancer research - an evaluation study. PLOS ONE, 14(12):1-21, 122019.
[9] Maisa Daoud and Michael Mayo. A survey of neural network-based cancer prediction models from microarray data. Artificial Intelligence in Medicine, 97, 062019.
[10] Nilufer Guler. Gene expression profiling in breast cancer and its effect on therapy selection in early-stage breast cancer. European Journal of Breast Health, 13:168-174, 092017.
[11] Joel Parker, Michael Mullins, Maggie Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John Quackenbush, Inge Stijleman, Juan Palazzo, J.S. Marron, Andrew Nobel, Elaine Mardis, Torsten Nielsen, Matthew Ellis, Charles Perou, and Philip Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 27:1160-7, 032009.
[12] Danielle Fitzpatrick, Christine Lai, Robert Parkyn, David Walters, Vladimir Humeniuk, and David Walsh. Time to breast cancer relapse predicted by primary tumour characteristics, not lymph node involvement. World journal of surgery, 38, 122013.
[13] Chenjing Zhu, Otilia Menyhart, Balázs Györffy, and Xia He. The prognostic association of spag5 gene expression in breast cancer patients with systematic therapy. BMC Cancer, 19, 112019.
[14] Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. CoRR, abs/1106.1813, 2011.
[15] Charles Perou, Therese Sørlie, Michael Eisen, Matt Rijn, Stefanie Jeffrey, Christian Rees, Jonathan Pollack, Douglas Ross, Hilde Johnsen, Lars Akslen, Øystein Fluge, Alexander Pergamenschikov, Cheryl Williams, Shirley Zhu, Per Lønning, Anne-Lise Børresen-Dale, Patrick Brown, and David Botstein. Molecular portraits of human breast tumours. Nature, 406:747-52, 092000.
[16] Therese Sørlie, Charles Perou, Rob Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael Eisen, Matt Rijn, Stefanie Jeffrey, Thor Thorsen, Hanne Quist, John Matese, Patrick Brown, David Botstein, Per Lønning, and A.L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implication. Proceedings of the National Academy of Sciences of the United States of America, 98:10869-74, 102001.
[17] Zhiyuan Hu, Cheng Fan, Daniel Oh, J Marron, Xiaping He, Bahjat Qaqish, Chad Livasy, Lisa Carey, Evangeline Reynolds, Lynn Dressler, Andrew Nobel, Joel Parker, Matthew Ewend, Lynda Sawyer, Junyuan Wu, Yudong Liu, Rita Nanda, Maria Tretiakova, Alejandra Orrico, and Charles Perou. The molecular portraits of breast tumors are conserved across microarray platforms. BMC genomics, 7:96, 022006.
[18] Cheng Fan, Daniel Oh, Lodewyk Wessels, Britta Weigelt, Dimitry Nuyten, Andrew Nobel, Laura van 't Veer, and Charles Perou. Concordance among gene-expression-based predictors for breast cancer. The New England journal of medicine, 355:560-9, 092006.
[19] Harold Hotelling. Relations between two sets of variates. Biometrika, 28(3/4):321377, 1936.
[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. Journal of Machine Learning Research, 9(86):2579-2605, 2008.
[21] Hege G. Russnes, Ole Christian Lingjærde, Anne-Lise Børresen-Dale, and Carlos Caldas. Breast cancer molecular stratification: From intrinsic subtypes to integrative clusters. The American Journal of Pathology, 187(10):2152-2162, 2017.
[22] Xiaofeng Dai, Ting Li, Zhonghu Bai, Yankun Yang, Xiuxia Liu, Jinling Zhan, and Bozhi Shi. Breast cancer intrinsic subtype classification, clinical use and future trends. American journal of cancer research, 5:2929-43, 122015.
[23] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.
[24] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. CoRR, abs/1705.07874, 2017.
[25] Christoph Molnar. Interpretable Machine Learning. 2019. https://christophm. github.io/interpretable-ml-book/.
[26] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282):457-481, 1958.
[27] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. Journal of the Royal Statistical Society. Series A (General), 135(2):185-207, 1972.
[28] Fabiana Tonellotto, Anke Bergmann, Karen Abrahão, Suzana Aguiar, and Marcelo Bello. Impact of number of positive lymph nodes and lymph node ratio on survival of women with node-positive breast cancer. European Journal of Breast Health, 15:7684, 042019.
[29] Samane Nematolahi and Seyyed Ayatollahi. A comparison of breast cancer survivals among young, middle and elderly aged patients of southern iran using cox and empirical bayesian additive hazard models. Epidemiology and Health, 39:e2017043, 10 2017.

## Appendices

| Variable | beta | p-value |
| :---: | :---: | :---: |
| lymph_nodes_positive | $6.878 \mathrm{e}+00$ | 0.009151 |
| NPI | $1.972 \mathrm{e}+00$ | 0.090502 |
| stage | $3.560 \mathrm{e}+00$ | 0.340729 |
| PROLIF | $2.141 \mathrm{e}+01$ | 0.078492 |
| ERBB2 | $3.486 \mathrm{e}+01$ | 0.147558 |
| age_at_diagnosis | $8.514 \mathrm{e}-02$ | 0.245331 |
| lymph_nodes_positive*NPI | $-1.471 \mathrm{e}+00$ | 0.004927 |
| lymph_nodes_positive*stage | $-4.191 \mathrm{e}+00$ | 0.001937 |
| NPI*stage | -9.483e-01 | 0.276196 |
| lymph_nodes_positive*PROLIF | $-1.926 \mathrm{e}+01$ | 0.005793 |
| NPI*PROLIF | $-4.771 \mathrm{e}+00$ | 0.121152 |
| stage*PROLIF | $-3.268 \mathrm{e}+00$ | 0.674616 |
| lymph_nodes_positive*ERBB2 | $-4.634 \mathrm{e}+01$ | 0.002670 |
| NPI*ERBB2 | $-5.765 \mathrm{e}+00$ | 0.333461 |
| stage*ERBB2 | $-3.453 \mathrm{e}+01$ | 0.095011 |
| PROLIF*ERBB2 | $-1.296 \mathrm{e}+02$ | 0.124561 |
| lymph_nodes_positive*age_at_diagnosis | -8.160e-02 | 0.049876 |
| NPI*age_at_diagnosis | -1.888e-02 | 0.308022 |
| stage*age_at_diagnosis | -5.173e-02 | 0.369840 |
| PROLIF*age_at_diagnosis | -3.304e-01 | 0.083950 |
| ERBB2*age_at_diagnosis | -4.092e-01 | 0.312343 |
| lymph_nodes_positive*NPI*stage | $8.753 \mathrm{e}-01$ | 0.001224 |
| lymph_nodes_positive*NPI*PROLIF | $4.126 \mathrm{e}+00$ | 0.001909 |
| lymph_nodes_positive*stage*PROLIF | $9.468 \mathrm{e}+00$ | 0.009881 |
| NPI*stage*PROLIF | $1.007 \mathrm{e}+00$ | 0.601811 |
| lymph_nodes_positive*NPI*ERBB2 | $9.121 \mathrm{e}+00$ | 0.001388 |
| lymph_nodes_positive*stage*ERBB2 | $3.015 \mathrm{e}+01$ | 0.000106 |
| NPI*stage*ERBB2 | $6.790 \mathrm{e}+00$ | 0.151856 |
| lymph_nodes_positive*PROLIF*ERBB2 | $1.534 \mathrm{e}+02$ | 0.002983 |
| NPI*PROLIF*ERBB2 | $2.921 \mathrm{e}+01$ | 0.163665 |
| stage*PROLIF*ERBB2 | $2.664 \mathrm{e}+01$ | 0.627782 |
| lymph_nodes_positive*NPI*age_at_diagnosis | $1.758 \mathrm{e}-02$ | 0.032853 |
| lymph_nodes_positive*stage*age_at_diagnosis | $5.723 \mathrm{e}-02$ | 0.007668 |
| NPI*stage*age_at_diagnosis | $1.208 \mathrm{e}-02$ | 0.376131 |
| lymph_nodes_positive*PROLIF*age_at_diagnosis | $3.097 \mathrm{e}-01$ | 0.004077 |
| NPI*PROLIF*age_at_diagnosis | 7.387e-02 | 0.128563 |
| stage*PROLIF*age_at_diagnosis | $6.848 \mathrm{e}-02$ | 0.574055 |
| lymph_nodes_positive*ERBB2*age_at_diagnosis | $6.301 \mathrm{e}-01$ | 0.011717 |
| NPI*ERBB2*age_at_diagnosis | $5.664 \mathrm{e}-02$ | 0.577024 |
| stage*ERBB2*age_at_diagnosis | $5.061 \mathrm{e}-01$ | 0.126118 |
| PROLIF*ERBB2*age_at_diagnosis | $2.146 \mathrm{e}+00$ | 0.107648 |
| lymph_nodes_positive*NPI*stage*PROLIF | $-2.013 \mathrm{e}+00$ | 0.002975 |
| lymph_nodes_positive*NPI*stage*ERBB2 | $-5.771 \mathrm{e}+00$ | $6.96 \mathrm{e}-05$ |
| lymph_nodes_positive*NPI*PROLIF*ERBB2 | $-3.228 \mathrm{e}+01$ | 0.000722 |
| lymph_nodes_positive*stage*PROLIF*ERBB2 | $-7.767 \mathrm{e}+01$ | 0.004222 |
| NPI*stage*PROLIF*ERBB2 | $-6.383 \mathrm{e}+00$ | 0.625433 |
| lymph_nodes_positive*NPI*stage*age_at_diagnosis | -1.174e-02 | 0.005975 |
| lymph_nodes_positive*NPI*PROLIF*age_at_diagnosis | $-6.454 \mathrm{e}-02$ | 0.001917 |
| lymph_nodes_positive*stage*PROLIF*age_at_diagnosis | -1.594e-01 | 0.004205 |
| NPI*stage*PROLIF*age_at_diagnosis | -1.618e-02 | 0.593267 |
| lymph_nodes_positive*NPI*ERBB2*age_at_diagnosis | -1.201e-01 | 0.009480 |
| lymph_nodes_positive*stage*ERBB2*age_at_diagnosis | -4.468e-01 | 0.000373 |
| NPI*stage*ERBB2*age_at_diagnosis | -8.933e-02 | 0.247849 |
| lymph_nodes_positive*PROLIF*ERBB2*age_at_diagnosis | $-2.345 \mathrm{e}+00$ | 0.003326 |
| NPI*PROLIF*ERBB2*age_at_diagnosis | -4.881e-01 | 0.144862 |
| stage*PROLIF*ERBB2*age_at_diagnosis | -5.320e-01 | 0.540693 |
| lymph_nodes_positive*NPI*stage*PROLIF*ERBB2 | $1.601 \mathrm{e}+01$ | 0.001058 |
| lymph_nodes_positive*NPI*stage*PROLIF*age_at_diagnosis | $3.268 \mathrm{e}-02$ | 0.001866 |
| lymph_nodes_positive*NPI*stage*ERBB2*age_at_diagnosis | $8.307 \mathrm{e}-02$ | 0.000360 |
| lymph_nodes_positive*NPI*PROLIF*ERBB2*age_at_diagnosis | $4.893 \mathrm{e}-01$ | 0.001061 |
| lymph_nodes_positive*stage*PROLIF*ERBB2*age_at_diagnosis | $1.241 \mathrm{e}+00$ | 0.002736 |
| NPI*stage*PROLIF*ERBB2*age_at_diagnosis | $1.157 \mathrm{e}-01$ | 0.580115 |
| lymph_nodes_positive*NPI*stage*PROLIF*ERBB2:age_at_diagnosis -2.519e-01 | 7.773e-01 | 0.000890 |

Table 16: Multivariate Cox regressions with all the patients.

| \# | lr | epochs | b._size | drop. | eps | L2-p | Tr | Val | Basal_val | Her2_val | LumA_val | LumB_val |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 0.01 | 80 | 100 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 97.7128 | 88.5106 | 82.3009 | 64.3836 | 85.7759 | 70.0599 |
| 2 | 0.01 | 80 | 100 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 94.6809 | 88.3511 | 86.1111 | 64.8649 | 91.7808 | 70.6522 |
| 3 | 0.01 | 80 | 150 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 96.8617 | 87.0213 | 84.0708 | 62.6506 | 86.6953 | 75.0 |
| 4 | 0.01 | 80 | 150 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 94.8404 | 88.4574 | 81.6667 | 65.2778 | 91.9431 | 70.3297 |
| 5 | 0.01 | 80 | 200 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 95.6383 | 87.9787 | 83.6207 | 72.1311 | 85.259 | 77.707 |
| 6 | 0.01 | 80 | 200 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 93.9894 | 88.1915 | 82.3009 | 70.5882 | 86.7769 | 71.6049 |
| 7 | 0.01 | 100 | 100 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 98.4043 | 88.2447 | 84.8214 | 62.6667 | 83.4025 | 71.3376 |
| 8 | 0.01 | 100 | 100 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 96.9681 | 89.5745 | 84.0708 | 65.2778 | 86.5801 | 72.1893 |
| 9 | 0.01 | 100 | 150 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 97.766 | 88.883 | 79.0323 | 60.4938 | 87.7193 | 72.3684 |
| 10 | 0.01 | 100 | 150 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 96.5957 | 87.4468 | 80.3419 | 62.5 | 88.2609 | 71.8391 |
| 11 | 0.01 | 100 | 200 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 95.0 | 88.4574 | 86.9159 | 62.5 | 86.6397 | 78.1457 |
| 12 | 0.01 | 100 | 200 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 97.9787 | 88.1383 | 81.8966 | 63.3803 | 83.6 | 74.3243 |
| 13 | 0.01 | 120 | 100 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 98.2979 | 90.1064 | 82.3529 | 62.8205 | 85.5372 | 76.7123 |
| 14 | 0.01 | 120 | 100 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 97.0745 | 88.5106 | 87.7358 | 62.3529 | 84.4262 | 76.0 |
| 15 | 0.01 | 120 | 150 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 98.9362 | 87.766 | 78.4 | 59.1549 | 86.9565 | 71.0692 |
| 16 | 0.01 | 120 | 150 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 96.5426 | 87.9255 | 90.099 | 64.3678 | 87.6596 | 77.1605 |
| 17 | 0.01 | 120 | 200 | 0.1 | $1 \mathrm{e}-08$ | 0.0001 | 97.2872 | 89.2553 | 80.1653 | 56.3218 | 88.0 | 72.3684 |
| 18 | 0.01 | 120 | 200 | 0.2 | $1 \mathrm{e}-08$ | 0.0001 | 97.7128 | 89.4149 | 81.8966 | 65.6716 | 87.069 | 71.7647 |
| 19 | 0.01 | 80 | 100 | 0.1 | $1 \mathrm{e}-08$ | 0.01 | 98.2979 | 82.3936 | 82.1429 | 66.1765 | 84.6154 | 71.519 |
| 20 | 0.01 | 80 | 100 | 0.2 | $1 \mathrm{e}-08$ | 0.01 | 94.734 | 83.2979 | 76.4228 | 57.3171 | 90.5172 | 76.3514 |
| 21 | 0.01 | 80 | 150 | 0.1 | $1 \mathrm{e}-08$ | 0.01 | 97.8191 | 69.5745 | 84.3478 | 64.1026 | 84.0 | 75.3521 |
| 22 | 0.01 | 80 | 150 | 0.2 | $1 \mathrm{e}-08$ | 0.01 | 95.6915 | 79.7872 | 83.0357 | 58.0645 | 87.5536 | 76.1905 |
| 23 | 0.01 | 80 | 200 | 0.1 | $1 \mathrm{e}-08$ | 0.01 | 95.5319 | 82.4468 | 81.5126 | 62.8205 | 88.8889 | 71.1656 |
| 24 | 0.01 | 80 | 200 | 0.2 | $1 \mathrm{e}-08$ | 0.01 | 95.4255 | 83.2979 | 86.3636 | 70.5882 | 87.0833 | 74.2515 |
| 25 | 0.01 | 100 | 100 | 0.1 | $1 \mathrm{e}-08$ | 0.01 | 97.8723 | 81.8085 | 78.8618 | 58.6667 | 88.5593 | 73.5099 |
| 26 | 0.01 | 100 | 100 | 0.2 | $1 \mathrm{e}-08$ | 0.01 | 98.2979 | 82.2872 | 84.0708 | 62.5 | 84.1004 | 73.2026 |
| 27 | 0.01 | 100 | 150 | 0.1 | $1 \mathrm{e}-08$ | 0.01 | 96.0106 | 83.5106 | 86.6667 | 69.8413 | 86.4629 | 68.617 |
| 28 | 0.01 | 100 | 150 | 0.2 | $1 \mathrm{e}-08$ | 0.01 | 97.0745 | 81.0106 | 86.4865 | 61.6438 | 87.4477 | 75.3086 |
| 29 | 0.01 | 100 | 200 | 0.1 | $1 \mathrm{e}-08$ | 0.01 | 96.2234 | 83.0851 | 80.3419 | 62.1622 | 89.9123 | 72.2892 |

71.4286
71.9745
72.028
73.2484
77.9221
74.1935
70.8075
71.1656
73.5099
76.8707
75.0
76.129
72.6619
74.8252
73.2919
69.3642
75.0
72.2973
73.3728
74.359
74.3421
67.4555
69.375
73.1544
73.4568
71.4286
69.7531
71.0227
74.8428
76.3889
77.2059












75.6944
77.1429
77.8571
73.9394
77.2727
73.6527
69.2308
74.0741
69.5946
73.9645
71.7105
75.7576
78.2609
75.153
76.4331
78.2895
79.0541
75.9036
78.4431
75.3165
80.1418
75.625
79.3103
75.6944
76.25
73.4568
77.551
75.9036
79.085
76.2821
77.551












76.5823
78.5714
77.027
80.0
75.974
77.7027
78.0
76.1589
77.8523
77.7778
76.7296
76.9737
75.6098
76.0736
77.2727
75.3247
76.5823
76.3975
75.3012
78.2895
76.6667
76.8707
77.2059
79.1139
75.641
76.129
76.6667
77.3585
77.4648
79.8658
76.129






 $\underset{0}{\circ} \boldsymbol{j}$ N





$$
\begin{aligned}
& 74.0 \\
& 75.9494 \\
& 78.8732 \\
& 75.3333 \\
& 76.5101 \\
& 77.4194 \\
& 77.0701 \\
& 74.0506 \\
& 75.7353 \\
& 79.3103 \\
& 76.5101 \\
& 75.3012 \\
& 79.7297 \\
& 75.1553 \\
& 77.1605 \\
& 72.9032 \\
& 78.7671 \\
& 76.8293 \\
& 77.0186 \\
& 76.129 \\
& 78.2313 \\
& 71.4286
\end{aligned}
$$

F N゙ N゙

 бু ৰু


Table 17: Accuracy results for cancer subtype prediction


[^0]:    ${ }^{1}$ SPCOM website: https://futur.upc.edu/SPCOM
    ${ }^{2}$ CALCULA services website: https://www.tsc.upc.edu/en/it-services/computing-services
    ${ }^{3}$ IDIBELL website: https://idibell.cat/

[^1]:    ${ }^{4}$ La Vanguardia [01/02/2021]: La covid retrasa el diagnóstico de cáncer de unas 55.000 personas.

[^2]:    ${ }^{5}$ The defining elements that distinguish one target from another.

[^3]:    ${ }^{6}$ Gene Set Enrichment Analysis. There is an R implementation in GSVA package

[^4]:    ${ }^{7}$ See the implementation in Stanford cibersortx webpage: https://cibersortx.stanford.edu/

[^5]:    ${ }^{8}$ KNNImputer package: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNN Imputer.html

[^6]:    ${ }^{9}$ Starts from an individual position to address the global variables.
    ${ }^{10}$ Arrangement of the clusters produced by the correlation computation

[^7]:    ${ }^{11}$ Adam optimizer implementation

[^8]:    ${ }^{12}$ A short course on Survival Analysis applied to the Financial Industry: Estimating survival by means of the Kaplan Meier estimator

[^9]:    ${ }^{13}$ https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-grades.html
    ${ }^{14}$ https://www.cancer.ca/en/cancer-information/cancer-type/breast/staging

