# SURVIVAL PREDICTION USING MULTIPLE INSTANCE LEARNING FOR TRIPLE NEGATIVE BREAST CANCER (TNBC)

Maria Alejandra Ortiz - Juan Sebastián Malagón - Juan Sebastián Neiza Mejía
Universidad Nacional de Colombia, Bogotá D.C., Colombia

Explanation video: https://www.youtube.com/watch?v=1CGCJkmH0Lw

## INTRODUCTION

Breast cancer is the most common malignancy in women. According to the WHO (World Health Organization), there were 2.3 million women diagnosed with breast cancer and 685.000 death globally in 2020.

There are two types of this cancer, that they have precision treatments. But there is one called the TNBC, due to its special molecular phenotype, that is not sensitive to endocrine therapy or molecular targeted therapy. For this reason, research has tried to improve the tissue description of this type of cancer.

The dataset used, has censored and uncensored data. A patient is censored when the time of entering and/or quitting to the clinical study is unknown; a patient is uncensored when this time information is known.

The dataset contains 128 censored and 23 non-censored patients. There is a big amount of censored data gathered around the median. This is due to the fact that there are a lot of patients with the same survival time and this can be noise for the model in posterior steps

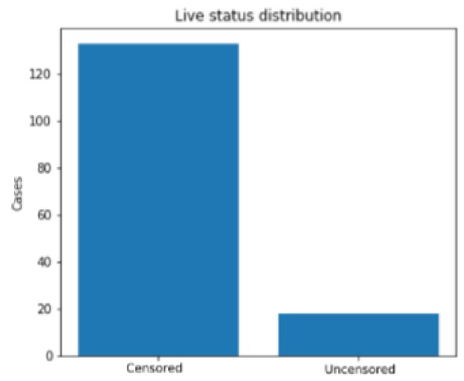



Figure 1: Live status distribution

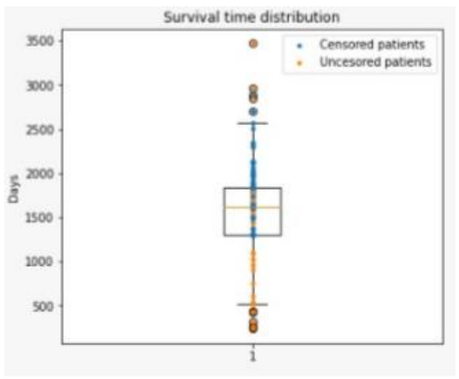
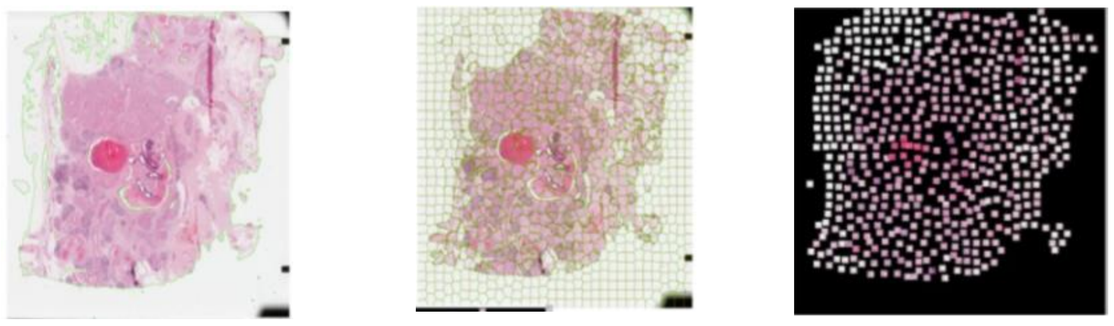



Figure 4: Survival time after imputation

## METHODOLOGY

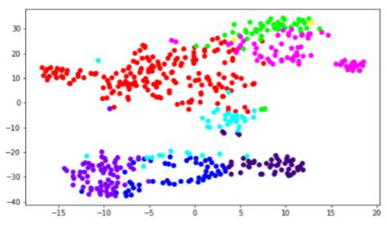
### PATCH CREATION BASED ON SUPERPIXELS:



The number of removable patches in each image is large. For this reason, in order to reduce the number of patches without affecting the phenotypic variability represented by them, the image is divided into superpixels, taking one patch for each super pixel (300 on average)

### PHENOTYPE CLUSTERING:

Once the compressed representations of each patch have been obtained, it is necessary to group them in order to identify the different phenotypes of the image. The figure on the right shows the clusters that indicate the different phenotypes present in the image.



On another hand, figure 10 shows this phenotype clusters applied into the WSI image sample.

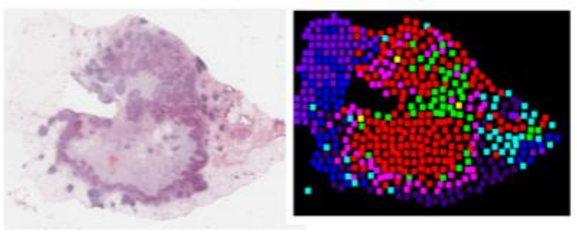


Figure 10: Phenotype clusters on WSI image sample

### SURVIVAL PREDICTION:

In order to predict the survival probability of a given patient, a **Multiple Instance Learning (MIL)** based architecture was built. This kind of model was chosen because the problem is weakly supervised and type of methods allows to learn from different representations of the images when labels are not present on them.

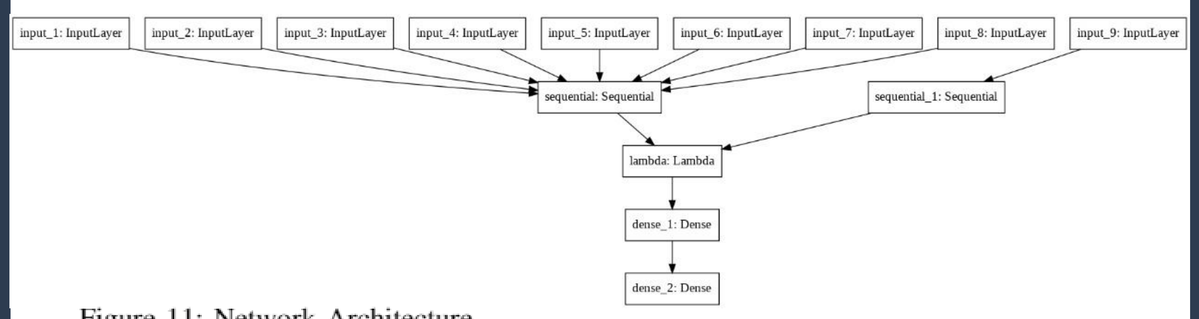
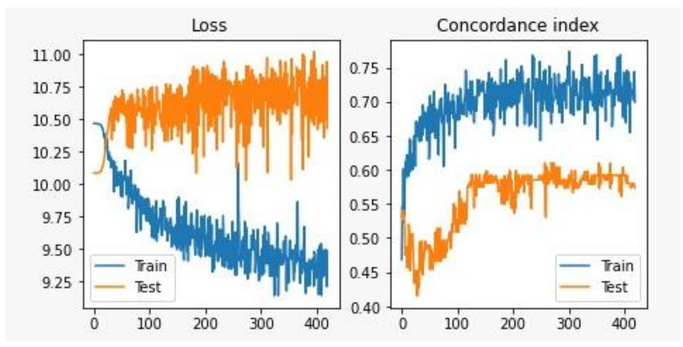



Figure 11: Network Architecture

## EXPERIMENTATION AND RESULTS

For the model proposed, the generated data after the preprocessing was divided for training, test and validation (70%, 20%and 10%, respectively). For the evaluation, a cross-validation process was implemented with 5 folds. The best combination of hyperparameters after the search were: Learning rate:0.001; Regularization factor: 0.00005; Bach size:8.

The values reached for the c-index and the valida-tion c-index for these combination of parameters were 0.6847 and 0.6232, respectively.




The behavior of the c index is the expectec with a value of 0,61 for the validation set. On another hand, the performance in the cost function indicates that there is overfitting, because the value in the cost function in the first epoch it tends to diverge as training progress.

| Method | c-index |
|---|---|
| DeepAtteMISL [8] | 0.69 |
| Lasso-Cox [10] | 0.4842 |
| Finetuned-WSISA-MTLSA [11] | 0.6428 |
| Proposed method | 0.61 |

Table I: C-index comparison

The table I compares the results above with the related works. Notice that the proposed model reached the values reported by previous works and seems to behave adequately with highly censored data.

## CONCLUSION

- The model proved to be robust in sets of highly censored data.
- Even with the noise added by the different processing techniques used, the model achieves results comparable to the state of the art.
- The model is scalable because it does not require markings provided by pathogens, only the wsi and the survival time.

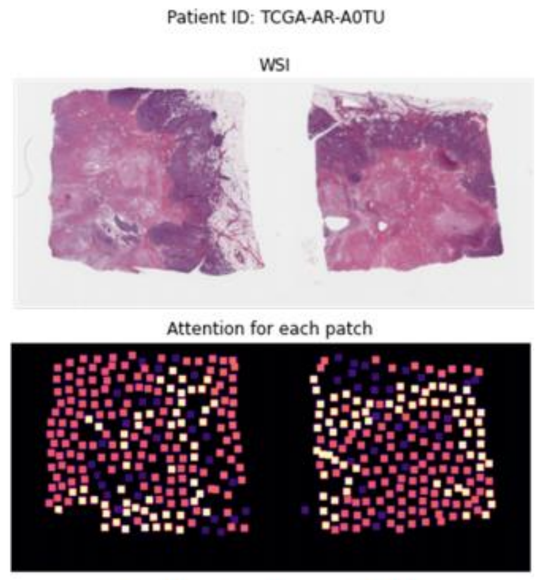



Figure 13: Patch attention on a WSI image

## FUTURE WORK

- It is recommended to use a stain normalization technique to improve the k-means and Bayesian methods to deal with the imputation of survival times.
- Hybrid methods could be implemented, such as Neural Random Forest, Deep Neural Kernel Methods, and so on.
- Explore different methods to calculate the weights based on instance representation.
- Another important point to take into account is to explore the proposed method using bigger optical amplifications which allow to obtain local features of each of the patch present in the images.