Optimisation of a Machine Learning algorithm applied to detection of cancer-associated fibroblasts

Author: Marc Nosàs Pomares
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.

Advisor: Núria Gavara Casas
(Dated: June 23, 2021)

Abstract: The activation of certain cells, such as fibroblasts, with the appearance of cancer has been studied and established. In order to accelerate and cheapen breast cancer diagnosis the optimisation of machine learning algorithms can be a powerful tool. Using two different data sets we studied the performance of different machine learning algorithms. We also studied the relevance of the different studied parameters, which could lead to some relevant biomedical conclusions. Further, we discuss the results obtained and discuss key aspects for improving the analysis of such experiments, as well as future directions for the use of Machine Learning in research against cancer.

I. INTRODUCTION

According to the "American Cancer Society" 1 in every 8 women suffer from breast cancer at some point of their lives [1]. World Health Organisation (WHO) stated that 2.3 million women were diagnosed with breast cancer and 685,000 died globally in 2020[2]. Breast cancer is the most prevalent in the world, and it affects people, specially women, at any age after puberty, even though the rates increase in older people.

Breast cancer arises in the glandular tissue, and it is initially innocuous. The uncontrolled growth of cancerous cells may progress and affect surrounding tissues and eventually be transmitted to blood vessels spread it throughout the body, affecting other organs. Women who die from breast cancer, do so because metastasis affects organs other than breast tissue, therefore identifying it in early stages and avoiding its spread is vital.

In effect, high-income countries where detection and treatment are more accessible to everyone the survival rates for at least 5 years after diagnosis are around 90%, whereas less developed countries have a survival rate of around 40%.

A. Cancer associated fibroblasts (CAFs)

Fibroblasts are non-specialised cells that form the connective tissue, present in most organs in the human body. Fibroblasts are normally non-active but they are capable of responding to tissue injury if activated; they play a main role in generating scar tissue. When normal fibroblasts are activated by tumour cells, they become either tumour-suppressing CAF or tumour-promoting CAF [3].

Fibroblasts also support the tissue function by regulating the extracellular matrix, by secreting fibrous proteins that form a supportive network reinforcing cell adhesion, proliferation and migration.

Numerous studies support that fibroblasts are associated with cancer cells at all the stages of cancer growth and play a main role for its spread or metastasis [4]. Therefore the presence of activated fibroblasts can be an indication to the presence of cancerous cells.

Among other functions tumour-promoting CAFs help remodel and regenerate cancerous tissue and support tumour growth, invasion of surrounding tissue and metastasis.

There are also some studies that support that, despite the evidence of the tumour-promoting role of the CAFs, fibroblasts are also involved in tumour suppression.

B. Machine Learning as a tool for research

Machine Learning (ML) relies on algorithms to analyse big datasets. Currently, Machine Learning has already earned a position in many fields, also in scientific research. ML is very useful as it is able to evolve and adapt new circumstances much faster than a human mind can. Algorithms detect patterns in a data source create behaviours based on the recognised patterns and make decisions based on the success or failure of its training. They work much faster than a human mind and the complexity of the things they can learn is expanding over time. Another important feature that makes ML a reliable tool is its cost, when an algorithm is created and the data set is gathered the cost of implementing it is very low compared to other research tools. So the main cost of machine learning is gathering enough data for the algorithm to be efficient, in the present study the data depends on the technique used to gather it, as different techniques have different costs.

The aim of this project is to implement a ML algorithm to study a set of CAFs and fibroblasts from healthy patients and train the algorithm to differentiate them. We are studying the cytoskeleton of the cells using two different techniques, both study a set of morphological parameters obtained by observing the actin protein of the cells with florescence [5]. In one case the cell sample has been obtained from people, affected by cancer, and people who

simply donated some breast cells due to some control or breast reduction. In the other case a sample of cells have been modified with TGF (transforming growth factor, [8]) drugs in order to activate the fibroblasts, artificially.

We believe that if we succeed on proving this algorithms useful, it should be very helpful for cancer diagnosis, as this algorithm should be able to study cells from apparently healthy patients and detect the probability of them suffering from, otherwise undetected, breast cancer. But it will also be useful to reach relevant biomedical conclusions, for instance which parameters are likely to be more relevant in the spread of cancerous cells and therefore which parameters should research highlight.

During the following sections the data used and the applied method are going to be discussed. And we aim to reach some relevant conclusions that help the scientific community in the fight against breast cancer.

II. THE DATA SET

During the study two different sets of data were used. One set of data was obtained by observing a group of 17004 single cells from 24 different people. The sample was given by the Breast Cancer Now Tissue Bank, [7]. The cytoskeleton of each cell was observed and parametrized with 14 parameters. In order to draw the structure of the cytoskeleton a chemical that reacted to the proteins of actin was used. For further information about this process I recommend the cited article, [5].

The other data set was studied the same way but the sample came from a group of artificially modified cells only from healthy patients. The size of this data set is considerably small, with only 3570 studied cells, from only 6 different people.

The data from the 24 patients was divided in 12 batches. Each batch contained cells of one healthy patients (a total of 12) and one patient that suffered breast cancer (a total of 12). From the patients that have suffered from cancer 4 different cancer types were studied, as listed in Table I. The age range studied in this study goes from 20 to 84 years old.

Depending on the location of the studied cell a 15th parameter was added, if the cell came from a healthy patient we considered a non-activated fibroblasts; if the cell came from an affected patient the cells were labelled with an S if it came from the tissue surrounding the tumour (although apparently healthy) or a T if it came from the affected tissue (therefore we considered this, activated fibroblasts). The number of cells in each group for each set of data is: 5807: N/ 5645: S/ 5552: T; actin fluorescence microscopy . And 1290: N/ 1161: S/ 1119: T; for the data obtained with TGF modified cells.

In order to standardise all the data, and evaluate it all with the same scale some modifications were performed before the study. A normalisation of every parameter

Batch	Cancer Type	Ages
1	Luminal A	28/84
2	Her2 enriched	52/54
3	Luminal B	25/48
4	Triple Negative	34/44
5	Her2 enriched	20/42
6	Luminal A	40/36
7	Luminal A	31/39
8	Luminal A	55/75
9	Lumnial A	48/47
10	Luminal A	57/47
11	Triple Negative	20/34
12	Triple Negative	65/48

TABLE I: Ages (healthy patient/cancer affected patient) and cancer types of each batch. Only the healthy patients from batches 7,8 and 9 were used in the study with tgf activated fibroblasts.

that set all the parameters between 0 and 1 using:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

After doing this normalisation we plotted an histogram for each parameter in each group (N/S/T), in order to observe if any of the parameters clearly drew a different behaviour for CAFs, none did.

In order to increase the efficiency of the algorithm we also decided to further standardise the sample using a z-score function, using:

$$x' = \frac{x - \mu}{\sigma} \tag{2}$$

Where x is the raw score, μ is the parameter mean, and σ is the standard deviation. It is a very useful standard score as it yields how far away the value is from the mean value.

A. Training and testing the algorithm

A key factor when it comes to machine learning is the set of data we use for the training. In this project the data was divided randomly,a random 5% of the data was used to test the results and the remaining 95% was used to train the algorithm. However, in order to avoid the randomness affecting the results, we have implemented the algorithm over 10 times and calculated the average results.

As there are 24 patients we were concerned that the correlation between cells of a same patient would affect the results. So that the algorithm would base its decision on detecting the patient, not the activation of fibroblasts. In order to ensure that wasn't the case we tried the algorithm using 11 batches as the training set and only one

batch to test the result, we did that for with every batch. We did observe a decrease in the accuracy of the algorithm, but it can be associated with the decline of the training size.

Even though we found that the correlation in a patient data is not relevant we observed something interesting when testing the algorithm with each batch. We observed that the accuracy among the healthy cells is similar (around a 70%), in every batch. However the accuracy in the affected tissue decreases a lot in some batches, probably this is related with the number of non-activated fibroblasts in the extirpated tissue.

III. THE ALGORITHM

With a proper data set we needed to find the most suitable algorithm to study it. We studied 3 algorithms that were likely to be suitable and tried to find which was the one with the best performance. We also observed how long it took for them to run and how variant they were. The three algorithms perform supervised learning, and are used for classification. Supervised learning occurs when a set of example data and associated target responses (in this case healthy cell or tumour associated cell) are studied by the algorithm to later predict the correct response when posed with new examples.

We used classification algorithms from the library Sckit Learn, in python. We considered three of the algorithms given of this library, tried them and timed their performance. The results obtained with the first data set are listed in TABLE II. The results obtained with the other data set are listed in TABLE III

Algorithm	Performance	Time (s)	CoV (%)
RFC (80)	0.73	0.51	1,4
Neural Network	0.69	1.07	2,0
SVC	0.70	0.31	1,8

TABLE II: List of all the algorithms that were tested, RFC (Random Forest Classification), Neural Networks and Support Vector Machines (SVM). The performance listed refers to the probability for the algorithm to reach the right answer. The number 80 in the parenthesis is the number of decision trees used to obtain this result, for further understanding see following sub section.

Algorithm	Performance	Time (s)	CoV (%)
RFC (74)	0.69	0.31	2,9
Neural Network	0.67	0.73	5
SVC	0.65	0.15	3,2

TABLE III: Results for the data set from artificially modified cells. The number 74 in the parenthesis is the number of decision trees used to obtain this result.

The Random Forest Classification (RFC) and Support Vector Machines (SVM) were considered the best to use for the study. Deciding to use Random Forest Classification in the end as the performance was slightly better and the size of the data set isn't long enough to notice the time difference. For future studies with larger datasets i would recommend to reconsider using an SVM algorithm, or any other algorithm that would present it self more efficient than RFC.

Notice that the results for the second data set are slightly less precise, however as the data set is much smaller the time inverted is shorter. Therefore this data can be considered better than the data obtained by fluorescence as it gives more information per cell. We believe that the difference between both data sets lies in the fact that in tumour affected tissue not all fibroblasts are activated, however in the artificially activated fibroblasts the percentage of CAFs is higher, and so it is the efficiency of training the algorithm. The last elements in the Tables II and III, referred to as Coefficient of Variance (CoV), is the standard deviation divided by the accuracy, it is an indicator of the spread of the results, and thus how much randomness affects the algorithm.

A. The Random Forest Classification algorithm:

The RFC is a classification algorithm based on decision trees. The algorithm generates N decision trees, each tree reaches a decision and classifies the input, then the algorithm uses averaging to improve the predictive accuracy. A number of statistical measurements determine how to make the splits in each decision tree. When we train the algorithm with a training dataset that has been already classified this splits are modified and grow the accuracy of the decision trees.

Obviously the more trees the algorithm uses the more accurate the result will be, but it will also take more time to run it. That is why we must study the accuracy of the algorithm versus the number of trees in order to optimise it. That is exactly what was done, in fig.1 we show the results obtained. As it can be seen in the figure the optimal number of trees is above 60 and below a 100, as above a 100 trees the accuracy of the algorithm plateaus and grows very slowly, making it is worthless to grow the number of trees anymore.

As it has been already stated there is a randomness associated with the results, but using 80 decision trees the accuracy is estimated around a 73% using the RFC algorithm.

IV. THE STUDY OF THE PARAMETERS

Another important point for the optimisations of the algorithm is to study what parameters are more important for the algorithm to classify data, we could see that some parameters are crucial and some are less relevant or even misleading. Furthermore studying the relevance of

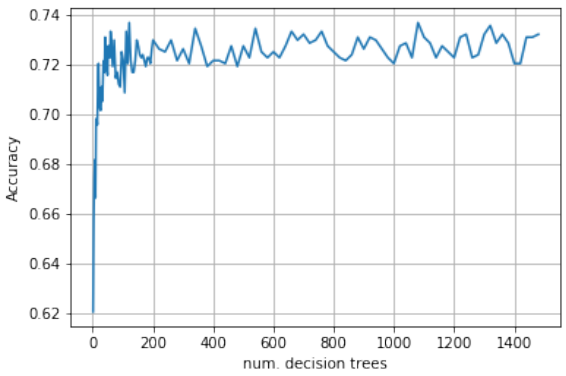

FIG. 1: Figure showing the accuracy of the algorithm vs the number of decision trees used to make the decision. This results concern the data obtained by using actin fluorescence microscopy.

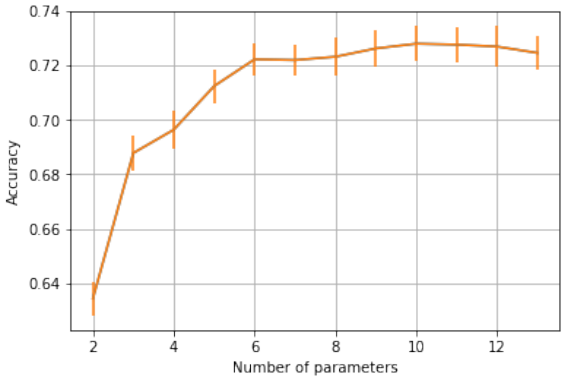

FIG. 2: In this figure we can observe the evolution of the accuracy of the parameters when taking of the most relevant parameters. Shown data: mean± SD, with N= 50 iterations

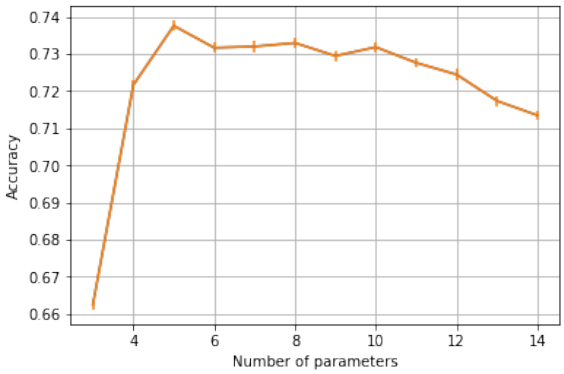

FIG. 3: Results for the artificially modified cells. Shown data: mean± SD, with N= 100 iterations.

the parameters for the classification might also rise some biomedical conclusions. Nevertheless it is important to take in mind that correlation and causation are not synonyms. That is why the results obtained in this section must be further studied in order to verify a causal relation between the relevant parameters and the activity of the fibroblasts.

In order to arrange the importance of each parameter we observed how the accuracy of the algorithm varied with and without each parameter. The parameter considered less relevant was removed from the matrix of parameters and the other parameters were further studied. We repeated this step 12 times, leaving only the 2 most important parameters left. We expected the curve to grow very fast at first, with only 2 parameters it is very difficult for it to achieve a good result. We expected this curve to grew very fast as the most important parameters were included in the data and to stop growing afterwards. Finally, some of the parameters might be misleading and so might cause a decrease in the accuracy.

In figures 2 and 3 the results obtained by this process can be observed.

The first figure refers to the results obtained with unmodified cells. The order of appearance of the parameters is as follows: 'Length variation','Alignment', 'Fibre length', 'Aspect ratio', 'Peak', 'Curvature', 'F-actin total amount', 'Spread','Thickness var', 'Stellate factor', 'Chirality', 'Chirality var'. The parameters that were left in the data set were: 'Area' and 'Fibre thickness'.

In the second case, concerning the TGF modified cells, the results were as follows: 'Alignment', 'Fibre thickness', 'Stellate factor', 'Length variation' , 'Chirality var', 'Curvature', 'Spread', 'Peak', 'F-actin total amount', 'Aspect ratio' ,'Thickness var', 'Chirality'. The parameters that were left in the data set were: 'Area' and 'Fibre length'.

We ran the code to create this figures more than once, and there was a high variation on the order of the last group of parameters, however, we could expect this variation we can not observe a change in the curve's behaviour with the removal of this parameters (they seem to be quite useless when it comes to decide whether or not the cell is affected), in the second figure we can observe even a decrease in the accuracy which would indicate that this parameters might be misleading.

V. CONCLUSIONS

With this study we were able to establish a correlation between a set of parameters of a group of fibroblasts and the presence of cancerous cells. We have written an algorithm capable of detecting this correlation and classifying this cells as cancer associated, although the accuracy of its classification could be improved in further studies. Moreover we observed that not all parameters are equally relevant, a further study on this parameters and their relation to cancerous cells, maybe by modifying this parameters with the use of drugs, could help curing the illness. Or rather stop it in earlier stages avoiding its growth and spread. Again, this must be further studied medically, the contribution of this study is showing the correlations between parameters and cancerous tissue.

The parameter that was demonstrated the most relevant was the Area of the cell's cytoskeleton, however some other parameters were relevant in both samples, for example the fibre morphology seams to affect the decision very much. Some other parameters do not have any relevance in the decision making, maybe we could remove them from further observations, it would make the code more efficient and we wouldn't have to assume the cost of measuring those parameters.

Given the results shown in the previous sections the probability of detecting CAFs as single cells is not high enough as to assure that the patient has cancer by only studying one cell, we consider a 73% of accuracy a low value. Nevertheless, as the studied set of data shows, we study many cells of every patient accordingly we encourage the study of samples of cells from every patient for an accurate diagnosis rather than the study of single cells. As we have already discussed the cells we study have to be within the tumour, as the surrounding tissue is not affected enough as to detect a relevant presence of CAFs.

The fact that we need a big sample of affected tissue is a little discouraging as it means that this algorithm is not a good tool to detect cancer in earlier stages, as we have to locate the tumour and extirpate some affected tissue to run the algorithms, without locating the algorithm we can not detect cancerous cells. However the idea of using ML to diagnose cancer can be used in other fields, such as day-to-day habits that are correlation to the apparition of cancer (such as smoking or doing sport), genetic history, or even a combination of all this different data sets. This algorithm can be useful once the cancer has been diagnosed, in order to confirm it, or given the case to detect what kind of cancer the patient is suffering. With a bigger data set this classification could be performed with a similar algorithm.

Concerning the study of cell groups we consider that trying other algorithms such as Support Vector Machines (SVM) would be very interesting, as despite the decrease in the accuracy, the efficiency of the algorithm would probably increase significantly. Besides adding data to the sample will rise the time of computing, thus a faster algorithm will be better to run a bigger data set, and the accuracy of the algorithm should improve significantly. Every data set has different properties and we must always find the most suitable algorithm for each data set, depending on its features and our propose.

Another possible further research concerns the fact that fibroblasts are not always cancer-promoting cells. Despite it is true that they play a main role in the growth and spread of the tumour, they can also be healing and cancer-suppressing. If we were able to compare this two groups of fibroblasts and differentiate them as we did between healthy and affected tissue, we might reach new conclusions. Knowing what parameters of an activated fibroblast we must affect to turn it from cancer promoting to cancer suppressing could be a key step to cure it.

During this study we haven't been able to create an effective tool for diagnosing cancer in very early stages, however we have created an algorithm that sets a seed in the direction of using what ML can offer in order to improve and make research more efficient. Furthermore, we have observed various parameters that are correlated with cancer and have pointed out which ones are more important to differentiate whether a cell is likely to be in a tumour affected tissue or not, this information is useful for further research on how to suppress CAFs by, maybe, modifying some of this parameters.

Acknowledgments

I could not finish this project without thanking everyone that has supported me through this last step on my degree. First and most I would like to thank my mentor, Núria Gavara. She has not only given me a lot of interesting bibliography but most importantly a lot of good advice and guidance. I must also thank Roser Sala, she has helped me and Núria with her experience in other physics final projects and also with her experience in the field of Machine Leaning.

Finally I also want to thank all my friends and family who have shown a lot of interest in the project and have filled me with a lot of good ideas.

[1] American Cancer Society. *Cancer Facts & Figures 2021.* Atlanta: American Cancer Society; 2021

[2] WHO: Breast Cancer, March 26th 2021, `https://www.who.int/news-room/fact-sheets/detail/breast-cancer`

[3] Gieniec, K.A., Butler, L.M., Worthley, D.L. et al. Cancer-associated fibroblasts—heroes or villains?. Br J Cancer 121, 293–302 (2019).

[4] Kalluri, R., Zeisberg, M. Fibroblasts in cancer. Nat Rev Cancer 6, 392–401 (2006).

[5] Flores, L.R., Keeling, M.C., Zhang, X. et al. Lifeact-TagGFP2 alters F-actin organization, cellular morphology and biophysical behaviour. Sci Rep 9, 3241 (2019).

[6] K.J Van Vliet, G Bao, S Suresh. The biomechanics toolbox: experimental approaches for living cells and biomolecules,Acta Materialia,Volume 51, Issue 19,2003,Pages 5881-5905

[7] https://breastcancernow.org/breast-cancer-research/breast-cancer-now-tissue-bank

[8] Brunen D, Willems SM, Kellner U, Midgley R, Simon I, Bernards R. TGF-: an emerging player in drug resistance. Cell Cycle. 2013 Sep 15;12(18):2960-8. doi: 10.4161/cc.26034. Epub 2013 Aug 12.