SEESLAB

Science and Engineering of Emerging Systems

# Cell cycle modelization through the study of the cell–cell and cell–matrix forces

Cristòfol Daudén Esmel

UNIVERSITAT ROVIRA i VIRGILI

**Rovira and Virgili University (URV)**

**Tutor: Marta Sales-Pardo**

17th June 2019, Tarragona

Final Degree Project

Biotechnology

# Index

# Investigation group

I have carried out my investigation project in the SEES Lab investigation group ( Science and Engineering of Emerging Systems ) in the Chemical Engineering Department, Rovira and Virgili University. Is a research group that is specialized in:

- **Complex Systems**, where individual components interact with each other, usually in non-linear ways, giving rise to complex networks of interactions that are neither totally regular nor totally random. Partly because of the interactions themselves and partly because of the interaction's topology, complex systems cannot be properly understood by just analyzing their constituent parts. Cells, ecosystems and economies are examples of complex systems.
- **Data Science,** Humans generate information at an unprecedented pace so processing this data requires new tools and new approaches at the interface of statistics, statistical and machine learning, network theory and statistical physics.
- **Multidisciplinarity**, the goal is to push forward the boundaries of science. They are interested in addressing fundamental questions in all areas of science including natural, social and economic sciences. Putting a special emphasis in the development of tools that aid scientific discovery through understanding and quantification of a specific phenomenon. To this end the group has assembled a multidisciplinary team and have established solid collaborations with experts in biology, social sciences, ecology and economics.

# Summary and keywords

The proliferation of cancerous tissues is directly related with the cell cycle speed and its duration. Control the cytological and physical-chemical factors that are involved in the phase change over the cell cycle and the cell division processes is necessary for the development of effective therapies for rapidly proliferating tumors.
Even though the regulation of the cell cycle has been a subject of extensive study for a long time, there are no still assessments of whether we can predict when a cell is going to divide or not. So, in this project, we have developed a statistical model that will allow us to predict when a cell is going to change its cell cycle phase, specifically, when the transition between the G1 phase and the S phase is going to happen.
The experiments carried out have generated some promising models that have quite good results on our dataset and also have shown the most promising factors in the prediction of this phase changing, such as the cumulative cytoplasmic area, the instantaneous tension and traction forces and the cumulative energies (calculated as the product between the cell area and the forces).

## Keywords

- **DNA ( deoxyribonucleic acid )**: polynucleotide formed from covalently linked deoxyribonucleotide units. It serves as the store of hereditary information within a cell and the carrier of this information from generation to generation.
- **S phase**: period of a eukaryotic cell cycle in which DNA is synthesized.
- **M phase**: period of the eukaryotic cell cycle during which the nucleus and cytoplasm divides.
- **G phase:** describes a cellular state outside of the replicative cell cycle.
- **Mitosis**: the division of the nucleus of a eukaryotic cell, involving condensation of the DNA into visible chromosomes, and separation of the duplicated chromosomes to form two identical sets. (From Greek *mitos,* a thread, referring to the threadlike appearance of the condensed chromosomes.).
- **Cytokinesis**: division of the cytoplasm of a plant or animal cell into two, as

distinct from the division of its nucleus (which is mitosis).

- **Overfitting:** the production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably. [Source: https://en.oxforddictionaries.com/definition/overfitting]
- **Selection bias:** error in choosing the individuals or groups to take part in a study. Ideally, the subjects in a study should be very similar to one another and to the larger population from which they are drawn. If there are important differences, the results of the study may not be valid. [Source: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/selection-bias?redirect=true]
- **Logistic function:** or **logistic curve** is a common "S" shape (sigmoid curve), with equation: $f(x)=\frac{L}{1+e^{-k(x-x_0)}}$
- **Sigmoid function:** mathematical function having a characteristic "S"-shaped curve or **sigmoid curve**.
- **Logit function:** or the **log-odds** is the logarithm of the odds $\frac{p}{1-p}$ where *p* is the probability. It is a type of function that creates a map of probability values from [ 0 , 1 ] to [ $-\infty$ , $+\infty$ ]. It is the inverse of the sigmoid "logistic" function.

# 1. Introduction

The cell cycle speed and duration is a very important factor in the proliferation of cancerous tissues. In order to develop effective therapies for rapidly proliferating tumors, it is necessary to control the physical-chemical and cytological factors that control these processes of phase changing over the cell cycle and cell division.

The regulation of the cell cycle has been a subject of extensive study for a long time. Intracellular pathways and soluble chemical factors have been hard analyzed to understand how they influence the cell cycle for decades (Saxton and Sabatini, 2017; Yu *et al.*, 2015; Lloyd, 2013). Early work, over the study of a single isolated cell, established that the shape and adhesion are potent regulators of DNA synthesis and cell growth (Folkman and Moscona, 1978, Watt *et al.*, 1988). However, it still remains unclear how the cell size (Chen *et al.*, 1997), the cell nucleus size (Roca-Cusachs *et al*., 2008), the growing rates (Son *et al*., 2012), the cytoskeletal tension (Huang *et al*., 1998) or the cell-ECM traction directly regulate the duration of the cell cycle.

Although these premises, it has been believed for a long time that cell division is regulated by the physical forces between cells and between the cell and the extracellular matrix (Mih *et al.*, 2012; Aragona *et al.,* 2013; Streichan *et al.,* 2014; LeGoff and Lecuit, 2015; Benham-Pyle *et al.,* 2015; Pinheiro *et al.,* 2017; Gudipaty *et al.* 2017; Lancaster, *et al*. 2013; Vianay *et al.* 2018). However, the evolution of these forces during the cell cycle in a tissue had never been measured, so it was unclear whether physical forces have an effect on the cell cycle progression.

Recently in a collaboration between the laboratory of Dr. X. Trepat (IBEC, Barcelona) and the Sees Lab - Rovira and Virgili University (Dr. Guimerà and Dr. Sales-Pardo), they made, for the first time, detailed measurements of the geometry of the cell and the forces that act on the cells belonging to a tissue during the cell cycle. These measurements have revealed mechanical-temporal patterns that regulate the duration of the cycle: the cells that undergo a greater tension are more likely to transit between the G1 and S phases and, in general, have a shorter cell cycle (G1 and S-G2-M phases).

At the moment, there is a model that explains the duration of the G1 phase with the accumulation of mechanical energy via mechanical stress (Uroz, M. *et al*. 2018) but there are no still assessments of whether we can predict when a cell is going to divide or not. So, this is the main objective of this project, to develop a statistical model that will allow us to predict when a cell is going to change its cell cycle phase, specifically, we want to predict when the transition between the G1 phase and the S phase is going to happen. To do that I use the magnitudes measured during the experiments in the paper from Uroz, M. *et al.*, such as the cell area, the tension forces and much more.

## 1.1. Cell Cycle

Cell reproduction begins with duplication of the cell's contents, followed by distribution of those contents into two daughter cells. Chromosome duplication occurs during the S phase of the cell cycle, whereas most other cell components are duplicated continuously throughout the cycle. During the M phase, the replicated chromosomes are segregated into individual nuclei (mitosis), and the cell then splits in two (cytokinesis). S phase and M phase are usually separated by gap phases called G1 and G2, where cell-cycle progression can be regulated by various intracellular and extracellular signals. Cell-cycle organization and control have been highly conserved during evolution, and studies in a wide range of systems - including yeasts, frog embryos, and mammalian cells in culture - have led to a unified view of eukaryotic cell-cycle control (Alberts *et al*., 2002).

## 1.2. G1/S transition

The transition from G1 phase, in which the cell grows, and the S phase, during which DNA is replicated, is a stage in the cell cycle at the boundary between these two phases. During this transition the cell makes decisions, based on its state, environmental cues and molecular signaling inputs, to become quiescent (enter G0), differentiate, make DNA repairs or proliferate. An accurate G1/S transition is crucial for the control of eukaryotic cell proliferation, and its misregulation promotes oncogenesis, so it is governed by cell cycle checkpoints to ensure the integrity of this

cycle. The subsequent S phase can also pause in response to improperly or partially replicated DNA (Bartek and Lukas, 2001; Bertoli *et al*., 2013),

# 2. Project hypotheses and objectives

Recent results in the literature show that intracellular tension and the cumulative mechanical energy are better predictors of the duration of the G1 phase than any other geometric property of the cell, such as the cell area or the growth rate of the cell area (Uroz, M. *et al*. 2018).

However, these studies have not attempted to make instantaneous predictions about whether a cell is going to divide or not in the near future taking into account its current physical/geometrical state/properties. In this project, I will work under the assumption that it is possible to accurately determine from physical properties when a cell will enter into cell division and the duration of the cycle.

The objective of the project is precisely to develop computational and mathematical models the use/take into account the measurements of geometric and physical properties of a cell to predict:

- the moment in which said cell enters cell division, when cell changes from the G1 phase to the S phase.
- the duration of the cell cycle.

# 3. Materials and Methods

The objective of the project is to develop a model that is able to predict when a cell enters cell division using some of the measured magnitudes of the given data, from now on I will refer to these magnitudes and combinations as features. To generate this model I had to do a preliminary exploration of the data, generate new features and get those features that are more significant for using them as an input for the model.

## 3.1. Data

The data I use in this project has been provided by Xavier Trepat's lab (Uroz, M. *et al*. 2018). This study reported for the first time measurements of the cell geometry and the forces that act on the cells in a tissue during the cell cycle. In addition to measuring the shape of all the cells, they also quantified cell–cell tension and cell–ECM (extracellular matrix) traction throughout the complete cycle of a large cell population in a growing epithelium. The process was done as follows:

> As a model system for epithelial growth, they used the expansion of a micropatterned colony of MDCK cells. They placed a polydimeth-ylsiloxane (PDMS) membrane with a 300-μ m-wide rectangular opening on top of a collagen-I-coated polyacrylamide gel (11 kPain stiffness) 26,27. To monitor the cell cycle during growth of the colony, they seeded MDCK-Fucci cells on the pattern and allowed them to adhere and form a confluent monolayer. MDCK-Fucci cells express Ctd1-red fluorescent protein (RFP) during G1 and S phases and geminin-green fluorescent protein (GFP) during the S–G2–M phases, which allowed them to monitor the state of each cell in the cycle and capture all the data.

The dataset consists of a set of 40 cells for which we have the following information over the entire cell cycle:

- cell size measured in $\mu m^2$
- cell nucleus size measured in $\mu m^2$
- cell-cell tension measured in $Pa$

- cell-ECM traction measured in $Pa$

From this initial data, I have generated some other features that we thought that will be useful during the analysis phase (see Figure 1). Some of them have been significant in existing results and the others come from that physical attributes that I think that can be affected during the evolution of the cell cycle:

- cell cytoplasmic size, which results in the difference between the cell size and the cell nucleus size.
- the mechanical energy that results of the product between the cell size and the cell-cell tension ($\mu J$), it will be called Energy from here onwards.
- the mechanical energy that results of the product between the cell size and the cell-ECM traction ($\mu J$), it will be called Energy 2 from here onwards.

The “Roll” Energies are calculated in the same way as the normal Energy but for each value I use a rolling window with the 5 last values.

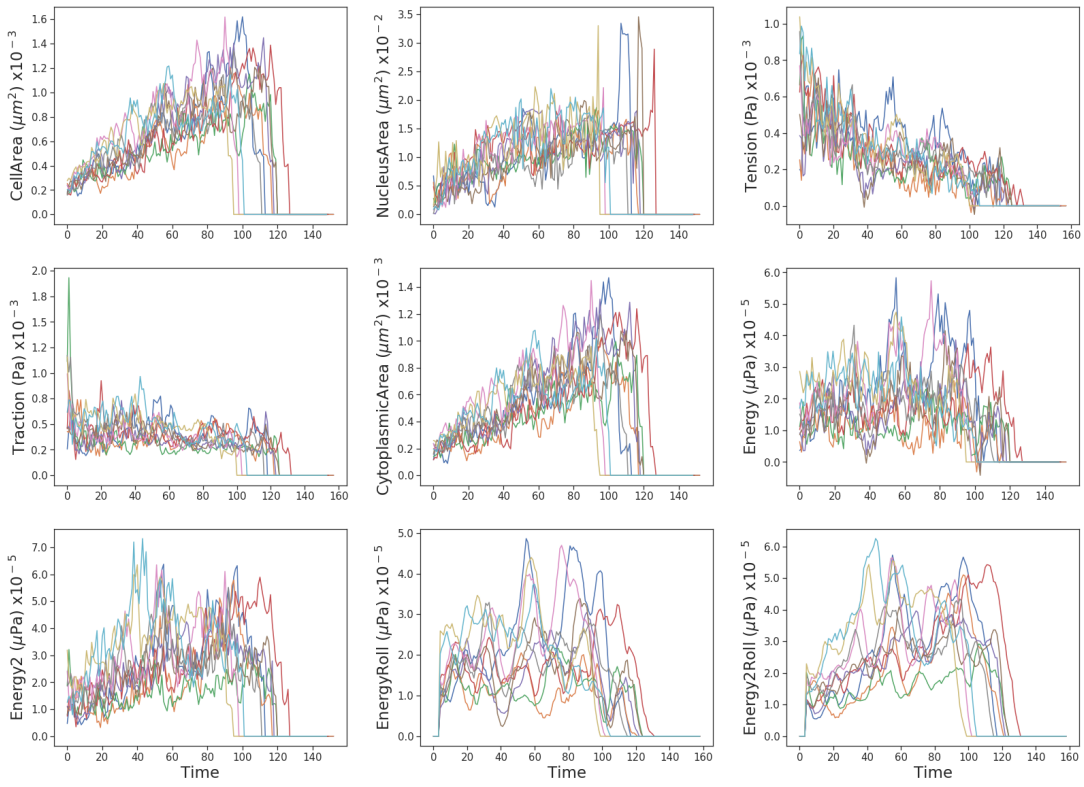


*Figure 1. Representation over the time of all the calculated features of 10 of the cells in the dataset. There is a moment from which the features get a 0 value, this is because, from that timestamp, there are no further measures. Energy corresponds to the product between the cell size and the cell-cell tension. Energy 2 corresponds to the product between the cell size and the cell-ECM traction. (full dataset in Appendix A Figure 18)*

### 3.1.1. Building the dataset

The first step was to build the dataset with which I had to work. The provided data is divided in 5 different csv format files. In the first file, the “divisions.txt” file, there is the information about the initial frame from which the measurements start, the frame in which happens the cell cycle transition from the G1 phase to the S phase and the frame in which the cell division happens (see an example in Figure 2).

| | Division_frame | x_f | y_f | Initial frame | S-G2-M |
|---|---|---|---|---|---|
| 0 | 114 | 7990 | 1095 | 1 | 68 |
| 1 | 118 | 8260 | 1170 | 1 | 76 |
| 2 | 124 | 8520 | 505 | 4 | 75 |
| 3 | 131 | 8305 | 435 | 4 | 87 |
| 4 | 123 | 10295 | 1280 | 5 | 82 |
| 5 | 125 | 10185 | 1375 | 5 | 84 |
| 6 | 103 | 9988 | 577 | 5 | 56 |
| 7 | 116 | 9868 | 403 | 5 | 59 |
| 8 | 100 | 10850 | 1245 | 5 | 55 |
| 9 | 106 | 10670 | 1405 | 5 | 64 |

*Figure 2. Representation of the data in the "divisions.txt" file of the 10 first cells. From each cell there is the information about the initial frame from which measurements are done, the G1 to S cell cycle phase change frame and the cell division frame.*

The other 4 files contain the information about the measures taken of the 4 measured features ( cell area, cell nucleus area and tension and traction forces ) in each time stamp. Each row of each file corresponds to all the measures taken of a single feature from a single cell (see an example in Figure 3).

| | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 182.4256 | 254.3360 | 168.4224 | 164.1216 | 238.8480 |
| 1 | 229.2480 | 221.8240 | 166.2720 | 168.5248 | 207.7952 |
| 2 | 233.8304 | 236.8000 | 203.5712 | 158.8992 | 218.3168 |
| 3 | 242.0992 | 267.2640 | 228.7360 | 230.2976 | 207.0272 |
| 4 | 195.2256 | 261.5808 | 185.9584 | 306.8928 | 280.6784 |
| 5 | 205.4656 | 236.1088 | 196.7360 | 224.8448 | 307.2512 |

*Figure 3. Content example of the "CArea.txt" file, in this table each column corresponds to a different cell, showing a total of 5 cells, and each row represents each time stamp in which the cell area measure was taken, showing a total amount of 6 time stamps.*

Once the data is loaded, I organize them into two different data frames (the Y and X variables). The Y variable has the information about if the cell is going to transite from the G1 phase to the S phase in the next 5 time stamps (True/False value) for each time stamp and cell (see an example in Figure 4). On the other hand, the X variable contains all the measured features values in each timestamp for all the cells, the feature values readen from the input files and the artificial generated features values, such as the cytoplasmic area, the energy values and the cumulative values (see an example in Figure 5).

| | cell | t | Switch |
|---|---|---|---|
| **0** | 0 | 4 | False |
| **1** | 0 | 5 | False |
| **2** | 0 | 6 | False |
| **3** | 0 | 7 | False |
| **4** | 0 | 8 | False |

*Figure 4. Y variable content example, in each row it is indicated the cell to which the data refers, the time stamp that we are considering and if in the next 5 time stamps the cell is going to change from the G1 to S cell cycle phase (here I am only showing the first 5 time stamps of the first cell).*

| | cell | t | CellArea | NucleusArea | Tension | Traction | CytoplasmicArea | Energy | Energy2 | accumulated__CellArea |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 4 | 195.2256 | 53.3248 | 657.46971 | 361.70620 | 141.9008 | 128354.918617 | 70614.309919 | 1082.8288 |
| **1** | 0 | 5 | 205.4656 | 49.8176 | 659.08398 | 353.67715 | 155.6480 | 135419.085401 | 72668.487831 | 1288.2944 |
| **2** | 0 | 6 | 150.9888 | 57.0112 | 429.38677 | 288.59586 | 93.9776 | 64832.593138 | 43574.742586 | 1439.2832 |
| **3** | 0 | 7 | 175.1808 | 61.6704 | 406.08491 | 347.66237 | 113.5104 | 71138.279402 | 60903.772106 | 1614.4640 |
| **4** | 0 | 8 | 169.4208 | 80.9984 | 434.49718 | 251.13199 | 88.4224 | 73612.859833 | 42546.982651 | 1783.8848 |

*5. Figure. X variable content example, in each row it is indicated the cell to which the data refers, the time stamp that we are considering and the values of all features with which I will work (because of the huge amount of data, here I am only showing the first 5 time stamps of the first cell and some of all the features).*

## 3.2. Cross-Validation

Cross-validation is an evaluation technique used to assess how the results obtained for the dataset being analyzed will generalize to an independent dataset. It is often used in predictive tasks, in which the goal is to estimate how the predictive model will perform in practice (e.g. out-of-sample predictions).

In a prediction problem, cross-validation works as follows. Suppose we have a dataset D:={($y_i$, $\mathbf{x}_i$)} where y is typically the dependent variable we want to predict and $\mathbf{x}_i$ is the vector of independent variables/features we want to use for prediction. Our goal is thus to develop a model/algorithm that predicts y from **x.** To assess the predictive power of a predictive model, we split the dataset into two parts: the training set and the test set. Then, first, we fit our model using the training set and, after that, we evaluate the trained model using the test set. Once the evaluation is done, we use the prediction results ($y_{pred}$) and the dependent variable real values (y) of the test set to evaluate the model performance.

There are two main types of cross-validation for the classification accuracy estimation (Joanneum, 2005-2006):

- K-fold Cross-validation: In this case, we generate an initial partition of D into K equal size subsets of a random permutation of the sample set, which are called folds. Then, we construct K train-test combinations. In each one of these, one fold is retained as the validation data for testing the model (test set) and the remaining K - 1 folds pooled together are used as training data (training set). Finally, we fit the model to the training set and its accuracy is evaluated on the test set (this process is repeated K times, once for each fold ). The resulting accuracy estimation comes from the average of all the accuracy estimations obtained after each iteration.
- Leave-one-out Cross-Validation: In this case the test set consists of only a point in the dataset and the remaining data composes the training set. As in the other case, for each test-train combination,we fit the model to the training set and we evaluate it with the test set (a single point, in this case).

Although these techniques are very useful in most of the cases, they are not suitable for our dataset. This is because our dataset consists of complete time-histories of

individual cells and therefore points in the trajectory of one cell are correlated with one another. As a result, for cross-validation purposes we need to consider the history of each individual cell as a single sample. We therefore designed a specific Cross-Validation type for our concrete use case, what we called Cell-fold Cross-Validation. It is quite similar to a  K-fold approach, but in this case each fold is not a random partition of the dataset, each fold corresponds to each one of the cells, doing as much iterations as cells are in the dataset.

For each cell-fold, we fit the model and make predictions for each one of the points in the test trajectory: ( 0, if the model predicts the cell will remain in the G1 phase during the 5 posterior time points, or 1, if the model predicts that the cell is going to change to the S phase).

### 3.2.1. Evaluation of cross-validation performance in classification tasks

A simple/way to visualize the performance of a model/algorithm in a cross-validation strategy is to build a confusion matrix (see Figure 6). A confusion matrix compares predicted classifications ($y_{pred}$ ) with real classifications (y). In this case, there are only two possible values for y=True (meaning that the cell will enter the S phase within the next five time steps) and False (meaning that the cell will remain in the actual cell cycle phase), therefore we can define:

- **True positives (TP):** the number of points for which $y_{pred}$=True and y=True. (equivalent to hit)
- **True negative (TN):** when a false sample is predicted as false (equivalent to correct rejection)
- **False positive (FP):** when a false sample is predicted as true (equivalent with false alarm)
- **False negative (FN):** when a true sample is predicted as false (equivalent with miss)

| **Confusion Matrix** | | Modeled Values: $y_{pred}$ | |
|---|---|---|---|
| | | False | True |
| Actual Values: y | False | True Negatives (TN) | False Positives (FP) |
| | True | False Negatives (FN) | True Positives (TP) |

*6. Figure: Confusion Matrix.*

## 3.3. Metrics

Once we had determined the evaluation technique that we were going to use to get the prediction results, we need to specify the metrics that will help us to understand these results. From the confusion matrix I get the precision and recall values:

- **Precision:** It is the proportion of predicted positive values that are a correctly real positive values.

$$Precision = \frac{true\ positive}{true\ positive + false\ positive}$$

- **Recall:** It is the fraction of positive values that have been retrieved over the total amount of positive values

$$Recall = \frac{true\ positive}{true\ positive + false\ negative}$$

From these two values, one can calculate the F1 Score which is a more realistic measure of our classifier's performance, it corrects those cases in which the arithmetic mean between the precision and recall can give us a wrong conclusion (like when the precision has a very high value and the recall is close to 0) (Sasaki, 2007; Chinchor, 1992):

- **F1 Score:** It is the harmonic mean of precision and recall.

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

  Note that its maximum value is 1 when Precision=Recall=1 and its lowest value is equal to 0 when either Precision or Recall are equal to 0.

When the models M we use to make predictions are generative models, that is the model assigns a probability to each prediction value $p(y_{pred}|M, x)$, we can estimate the plausibility of the overall predictions the model makes using the held-out likelihood.

- **Held-out likelihood:** It is the likelihood (or log-likelihood) of the test set given the model obtained by fitting the training set. If we assume that each observation is independent, the held-out log likelihood can be written as:

$$\mathscr{L}(X) = \log L(X) = \sum_{i=1}^{X} \log p(y_i|M, X_i)$$

  where $X$ is the test data set, $y$ is the class to which each sample belongs to and $M$ is the model (Wallach *et al.,* 2009). In general, this metric penalizes

overfitted models which will give high probabilities for the training set, but generalize poorly, so that probabilities for the test set are low. A higher likelihood value implies a more predictive model.

## 3.4. Machine Learning

Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. Typically, an example $x$ is represented by a tuple of attribute values $(x_1, x_2, \ldots, x_n)$, where $x_i$ is the value of attribute $i$. Let $Y$ represent the classification variable, and let $y$ be the value of $Y$.

### 3.4.1. Logistic Regression

In statistics, it is a model that uses a logistic function to model a binary dependent variable. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

The model dependent variable has two possible values, such as pass/file, win/loss, or in our case, the phase change happens or not, where these values are labeled “0” and “1”. The probability of each one of the two values is expressed in terms of a linear combination of one or more independent variables (“predictors”) which can be categorical (two classes) or continuous (any real value).

In logistic regression, $p(y=1|x)$ is modeled via the logistic function - a sigmoid function that takes any real input and outputs a value between zero and one. The logistic function $g(z)$ is defined as follows:

$$g(z)=\frac{e^z}{e^z+1}=\frac{1}{1+e^{-z}}$$

If we assume that $z$ is a linear combination of multiple explanatory variables $x_1, x_2, \ldots, x_n$. We can then express $z$ as follows:

$$z=\alpha_0+\alpha_1 x_1+\alpha_2 x_2+\ldots+\alpha_n x_n=\alpha_0+\sum_{i=1}^{n}\alpha_i x_i$$

So $p(y=1|x,\theta)$ can be written as:

$$p(y=1|x;\theta)=\frac{1}{1+e^{-(\alpha_0+\sum_{i=1}^{n}\alpha_1 x_1)}}$$

where θ are the parameters of the linear model ($\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n$).

Note that by consistency, $p(y=0|x;\theta)=1-p(y=1|x;\theta)$.

To train the model, I compute the error of the predictions with respect to the real values with Cost function (Maximum log-Likelihood)

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta(x^{(i)})-y^{(i)}\right)$$

$$A=\frac{-1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log\left(h_\theta(x^{(i)})\right)+\left(1-y^{(i)}\right)\log\left(1-h_\theta(x^{(i)})\right)\right]$$

and I adjust the initial $\theta$ values by applying the Gradient descent algorithm to minimize the cost function value.

$$\theta_j=\theta_j-\alpha\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)})-y^{(i)}\right)x_j^{(i)}$$

Now, the Logistic regression model is ready to predict new data. Using the new x values as an input we get the predicted y values.

### 3.4.2. Random Forest Classifier

Random Forest Classifier (RF) is an ensemble machine learning method for classification.

RF operates with decision trees. A decision tree is a representation of a decision procedure for determining the class ($y$) of a given instance ($x$). Each node of the tree represents a feature that partitions the space of instances at the node according to the possible outcomes of the test. Each subset of the partition corresponds to a classification subproblem for that subspace of the instance, which is solved by a subtree. So, each link (branch) represents a decision (rule) that will generate a new subtree to solve a part of the initial problem and each leaf represents an outcome (categorical or continuous value) (Utgoff, 1989).

During the training phase, RF operates by constructing an ensemble of decision trees, where each one of them will fit to a subset of the training dataset, this is what we call a forest. After, when using this model to do a prediction over a sample, the algorithm works as follows: each one of the decision trees makes a prediction, and

the forest prediction comes from the all individual trees prediction mode/average. The main difference between a normal decision tree, like the ID3 decision tree, and the ones used in this algorithm is that when considering the features to do the space partition on each node, it only considers a random subset of all the available features (from here comes the “Random” of the algorithm name).

This classifier fixes the overfitting problem that appears in decision trees. The RF is less sensible to little input data changes than the decision tree and has a better global precision in classification tasks. However, the computational cost in creating the relational forest and using it is higher and explaining the results obtained is harder than using a single decision tree.

### 3.4.3. Gaussian Naive Bayes

Naive Bayes (NB) is the simplest form of Bayesian network, in which we assume that all $p(x_1,\dots,x_n|y)=\prod_{i=1}^{n} p(x_i|y)$, i.e. attributes are independently affected by the class value y (Zhang, 2004).

As a conditional probability model, naive Bayes assigns a probability $p(y|x_1,\dots,x_n)$ for each of the possible classes to a problem instance. This instance is represented by a vector $x=(x_1,\dots,x_n)$ representing some $n$ features (independent variables). The problem with this formulation appears when the number of features $n$ is large or if a feature can take on a large number values, then basing such a model on probability tables is infeasible. To solve this issue we reformulate the model to make it more tractable by using Bayes' theorem, the conditional probability can be decomposed as:

$$p(y|x)=\frac{P(y)P(x|y)}{P(x)}=\frac{P(y)\prod_{i=1}^{n} p(x_i|y)}{P(x)},$$

which using Bayesian probability terminology can be written as:

$$posterior=\frac{prior\times likelihood}{evidence}$$

As we use Gaussian NB, the likelihood of the features is assumed to be Gaussian:

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(\frac{-(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$

The naive Bayes classifier combines this naive Bayes probability model with a decision rule. We use the MAP (maximum a posteriori) decision rule which consists in picking the hypothesis that is most probable. The corresponding Bayes classifier is a function that assigns a class label $\underline{y} = y$ for some $Y$ as follows:

$$\underline{y} = arg\,max_y\, p(y) \prod_{i=1}^{n} p(x_i|y)$$

# 4. Results and discussion

## 4.1. Preliminary Analysis

My first step was to represent the time evolution of the different measured quantities (features) for every cell. As you can see in Figure 1 the values for all the features from the different cells followed the same trend.

The cell area and the cell nucleus area increase over time while tension and force decrease. On the other hand, the traction force seems to remain stable. The other variables shown in the figure, correspond to features generated from the first ones (cell and nucleus areas and tension and traction forces). The cytoplasmic area and the energies calculated using the traction force increase over the time while the energies calculated using the tension force seems to remain stable.

A remarkable fact is that while the trends are clear, the values in each timestamp were quite different between the cells; this can become a problem when treating the data.

My next step was to analyze correlations between variables (Appendix B Figure 19), since highly correlated variables are redundant for modeling and predictive purposes. To my surprise, the cell area and the cell nucleus area were not as correlated as I would have expected, and the tension and traction forces are weakly correlated with the other features, so we can conclude that these features are independent from the other features. The cytoplasmic cell area is strong correlated to the cell area but not to the cell nucleus area, this makes sense because this

feature comes from the difference between these two features, having the cell area a much higher value than the cell nucleus area. Finally, the energies which are calculated using the tension force are more correlated to the cell area than the ones calculated using the traction force.

Next, my goal was to identify the proper features number to use in the predictive analysis. To that end, I picked up different sets of all the initial data (sets had a different number of features, and the sets with the same size had a different combination of those features), and I evaluated the performance of three models I consider (Logistic Regression, Random Forest Classifier and Naive Bayes), using the cell-fold cross validation method.

Here, I observed that using all the features for predictions does not necessarily entail getting the better results. When using a high number of characteristics seems that the algorithms have troubles in treating all the data and the results are worse than when I use a small number of features. So, as our objective is to create a mathematical model, we decided to use only two features to create an initial simple model and study how it works with the data.

## 4.2. Selection of the best machine learning algorithm and features for prediction

The objective in this phase was to assess which of the algorithms used had better prediction results and with which features. To do that, I examined all possible two-feature combinations with each algorithm. Then I picked up those combinations that had a better metrics results (superposing the Held-Out likelihood value over the other metrics).

### 4.2.1. Overall comparison of the algorithms performance

An initial held-out likelihood analysis of the used algorithms gave the results represented in the following figures:

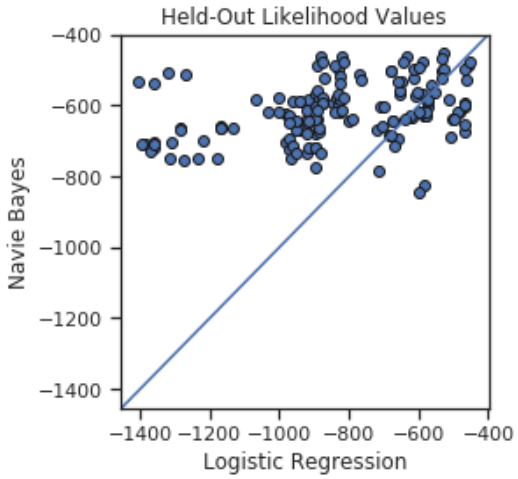


*Figure 7. Comparison of the Held-out likelihood values obtained in the evaluation of the Gaussian Naive Bayes and Logistic Regression models using as an input all possible two features combinations.*

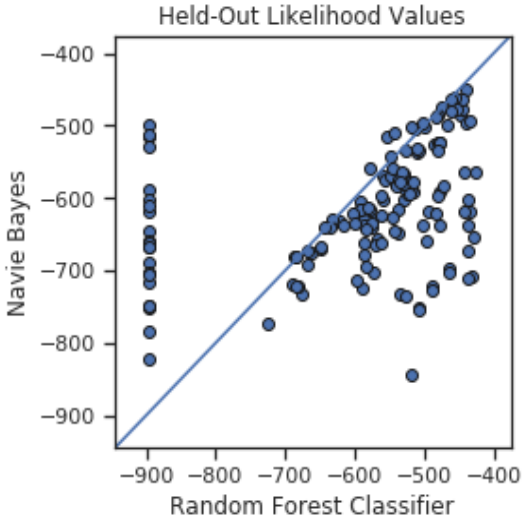


*Figure 8. Comparison of the Held-out likelihood values obtained in the evaluation of the Gaussian Naive Bayes and Random Forest Classifier models using as an input all possible two features combinations.*

Figure 7 shows that the Logistic regression algorithm has the worst performance. It has a lower held-out likelihood value for most of the feature combinations and the best F1 Score achieved is lower than 0.4, which is lower than the values achieved with the Gaussian Naive Bayes algorithm (Tables 11, 12, 15 and 16 Appendices C and E). On the other hand, Random Forest has a better held-out likelihood values than Gaussian NB in most of the cases. Unfortunately, there are some of them in which I get a minus infinite value (points at the left in Figure 8), this means that the

algorithm overfits the dataset. F1 scores (see Table 14 Appendix D and Table 16 Appendix E ) reinforce the selection of Gaussian NB as the best performing algorithm for this task, since F1 score values are higher than the ones achieved by the RF algorithm, the NB Gaussian model does not overfit the dataset and it also has good held-out likelihood values.

### 4.2.2. Analysis of the best models for each algorithm

In this section, I analyze the results obtained by the different algorithms in terms of performance and of the best biological features for prediction.

#### 4.2.2.1. Best Logistic Regression model results

The best result I obtain using Logistic Regression(LR) corresponds to using the traction force and the cumulative nucleus area as features (see Table 1). Using these two features, I get a quite good held-out likelihood (see Figs 7 and 8 for likelihood values) and a precision metric of 0.65. Unfortunately, the recall value is too low (I consider low values those which are under 0.35) to consider LR a good predictive model.

| Best 2 Features Logistic Regression Model Result | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Traction | cumulative Nucleus Area | -452.47 | 0.38 | 0.65 | 0.27 |

*Table 1. Best LR result.*

In fact, taking a look at global results (see Tables 11 and 12 Appendix C), it becomes apparent that LR priorizes precision over recall. This means that the model aims at minimizing false positives rather than maximizing the number of true positives. So, using a model like this one, you will detect less phase changes but that ones detected have a higher probability of not being a false positive.

Best models generated by this algorithm suggest that the best features for predicting the phase change are the traction force, the cumulative cell nucleus and the cytoplasmic areas and the mechanical energy that results of the product between the cell size and the cell-ECM traction (Energy 2).

#### 4.2.2.2. Best Random Forest Classifier results

Random Forest results are not as consistent as those of LR, therefore I am going to discuss three different scenarios.

First, if I consider the cumulative traction force and the cumulative cytoplasmic area as features, I get the best held-out likelihood value; in fact, this is the best held-out likelihood value obtained taking into account all three models and all possible feature combinations. Despite the result in this metric, the other results are under the 0.2 value, in particular, the recall value is 0.07 which means that this model is only able to predict the 7% of the cases in which the phase change is going to happen.

The next case is when using the cumulative cell area and the cumulative tension force. Using these features, in spite of simultaneously getting the higher recall and F1 Score, these results are not as good as those obtained with the Logistic Regression model.

The last scenario is when I use the Energy roll and the cumulative tension force as features. Here, I get the maximum precision possible but also the lowest recall, which means that although this model guesses correctly when predicting that a phase change is going to happen, it is only able to recover the 5% of all true positives.

| Best 2 Features Random Forest Classifier Results | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Traction | cumulative Cytoplasmic Area | -423.85 | 0.10 | 0.19 | 0.07 |
| cumulative Cell Area | cumulative Tension | -431.93 | 0.36 | 0.62 | 0.25 |
| Energy roll | cumulative Tension | -600.66 | 0.10 | 1.00 | 0.05 |

*Table 2. Best RF results, each row corresponds to the best result taking into account one of the metrics: held-out likelihood, F1 score and Precision respectively.*

Taking a look at the global results (see Tables 13 and 14 Appendix D), RF, as LR does, prioritizes high precision rather than high recall.

In this case, the best models generated by the RF algorithm suggest that the best features for predicting the phase change are the cumulative tension force, the cumulative cytoplasmic area and the Energy 2.

#### 4.2.2.3. Best Gaussian Naive Bayes model results

As in the case of the RF, I consider different cases.

The best held-out likelihood score is achieved when using the Traction force and the cumulative cell area as features (see Table 3), which is better than the best LR results (see Appendix C Table 11) but worse than the best RF result (-423.85).

The best F1 score values are achieved when using the cumulative nucleus area with the cumulative cytoplasmic area or with the cumulative cell area (Appendix E Table 16), but, as these pair of features are quite correlated between them, I also take into account the pair formed by the cumulative cytoplasmic area and cumulative Energy 2, which has the third best F1 score value.

Finally, the best precision value is achieved by using the feature pair of tension force and cumulative cell area and the best recall value by using the cumulative cell area and the cumulative nucleus area.

| 2 Features Gaussian Naive Bayes Model Results | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Traction | cumulative Cell Area | -449.62 | 0.43 | 0.45 | 0.42 |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -653.92 | 0.50 | 0.40 | 0.67 |
| cumulative Cytoplasmic Area | cumulative Energy 2 | -706.27 | 0.48 | 0.40 | 0.60 |
| Tension | cumulative Cell Area | -463.94 | 0.46 | 0.46 | 0.46 |
| cumulative Cell Area | cumulative Nucleus Area | -671.53 | 0.49 | 0.39 | 0.68 |

*Table 3. Best Gaussian NB results, each section corresponds to the best result taking into account one of the metrics: held-out likelihood, F1 score, Precision and Recall respectively.*

Taking a look to global results (see Tables 15 and 16 Appendix E), in contrast with the models generated with the other algorithms, the Gaussian NB models prioritizes a high recall over the precision, even though the precision values are not as low as the recall values achieved by the other models.

Best models generated by the NB algorithm suggest that the best features for predicting the phase change are the tension and traction forces, the cumulative cytoplasmic and cell areas and the cumulative Energy 2.

## 4.3. Analysis and Interpretation of the best models

As I have already discussed Gaussian NB is the overall best performing algorithm for the classification task. By merging the model results discussed in the previous sections, although their performances are different, the best features are consistent across the algorithms. So, the best pairs of feature combinations are the following:

- the accumulative cytoplasmic area and the cumulative Energy 2 (calculated as the product between the traction force and the cell area)
- the tension force and cumulative cell area
- the traction force and cumulative cell area

Now, to define a mathematical model first I wanted to identify trends in the predictive models. To do that, I had to visualize what the Gaussian NB looks like in terms of model predictions for different feature values. For each one of the best pairs of feature combinations, I created a grid of values of the two selected features, and then plotted the prediction of the trained model for each pair of values of the features.

As RF has the best held-out likelihood prediction results, I visualized the model in the same manner to understand which is the difference between the two algorithms and see if this one is also a good candidate to be represented by a mathematical function.

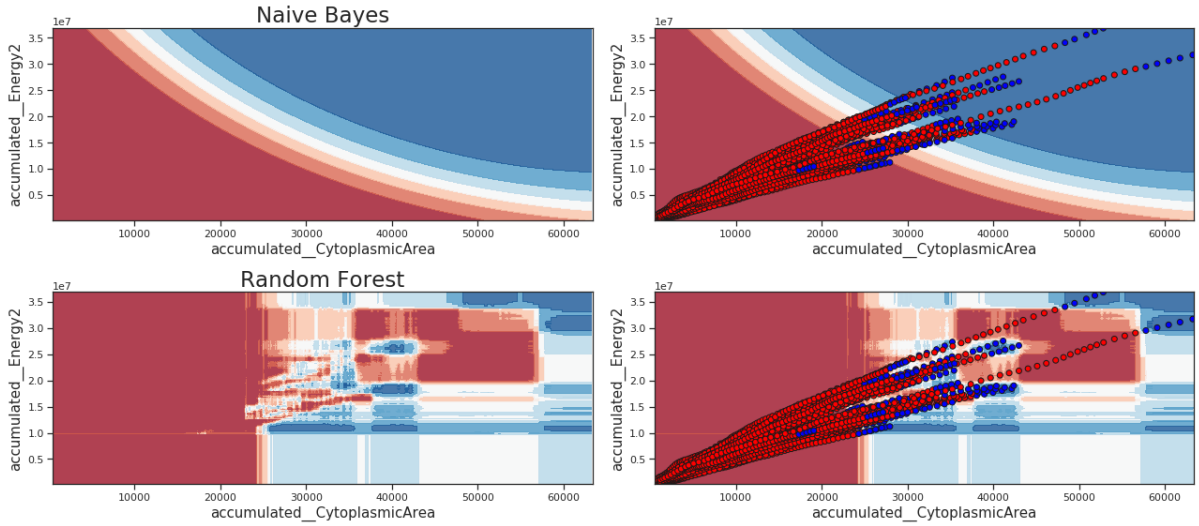


*Figure 9. Gaussian Naive Bayes and Random Forest models trained using the cumulative cytoplasmic area and the cumulative energy (the result of the product between the cell area and the traction force) features. In the first column it is shown models behaviour (the gradient colour indicates the prediction value, where probability 0 is represented by the colour red and 1 is represented by the colour blue). In the second column our dataset is added (red points mean that the cell is going to remain in the G1 cell cycle phase for the next 5 time stamps, while blue points mean that the cell is going to change from G1 to S phase in the next 5 time stamps).*

Figure 9 shows that RF Classifier overfits the training set, the model gets good results but only for this dataset, if I use some data that is quite different from the one I have used on the training phase, the results will have no sense. On the other hand, the NB model seems to perform bad predictions, as the model generated does not fit the data used in the training phase. Despite this, when I consider the metrics used to evaluate the model performance, the results (see Table 4) show that the Gaussian NB model does better predictions than the RF model.

| | Held-Out Likelihood | F1 Score | Precision | Recall |
|---|---|---|---|---|
| Gaussian Naive Bayes | -706.3 | 0.48 | 0.40 | 0.60 |
| Random Forest Classifier | -inf | 0.24 | 0.28 | 0.21 |

*Table 4. First comparison between the bests Gaussian NB and RF Classifier models.*

Table 4 shows that the held-out likelihood value reaches a minus infinity value on the random forest model, which means that the model overfits the dataset, and a -706.30 value on the NB. About the other metrics, in all of them, the NB results are 0.2 points over the ones obtained with the RF. The most significant result in this

analysis is that the NB model is able to recover, to predict properly, the 60% of the true values in the test set.

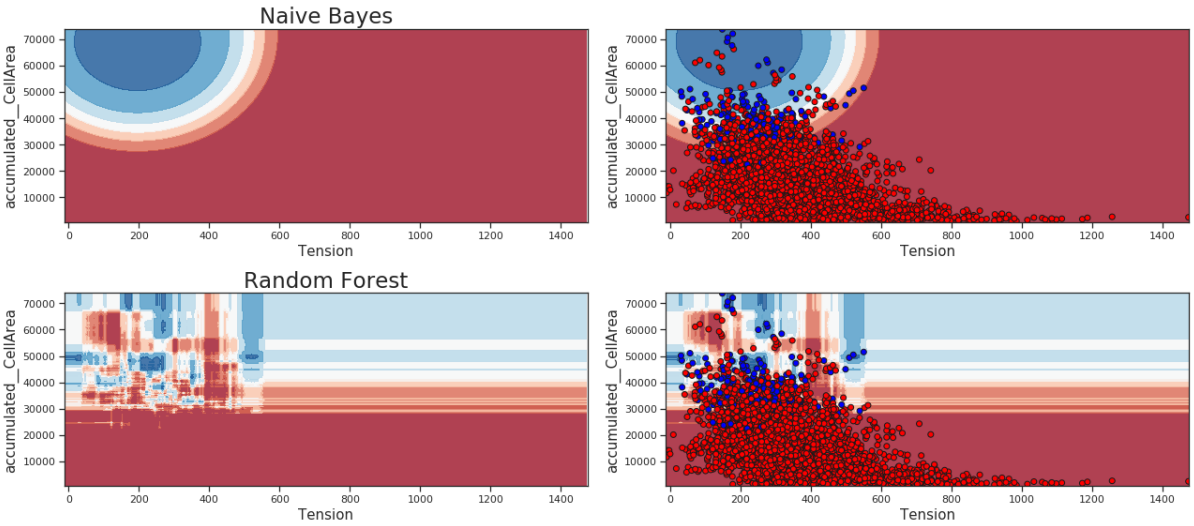


*Figure 10. Gaussian Naive Bayes and Random Forest models trained using the tension force and the cumulative cell area features. In the first column it is shown models behaviour and in the second column our dataset is added to the initial representation.*

In this case, happens the same than in the previous analysis, the RF Classifier model overfits the training set, as can be seen in Figure 10, and I get better results with the Gaussian NB model than with the other one (see Table 5).

| | Held-Out Likelihood | F1 Score | Precision | Recall |
|---|---|---|---|---|
| Gaussian Naive Bayes | -463.9 | 0.46 | 0.46 | 0.46 |
| Random Forest Classifier | -inf | 0.28 | 0.34 | 0.23 |

*Table 5. Second Comparison between the bests Gaussian NB and RF Classifier models.*

Results in Table 5 show that in spite of RF model results are higher than the ones obtained with the previous RF model, they still do not reach the ones achieved by the NB models.

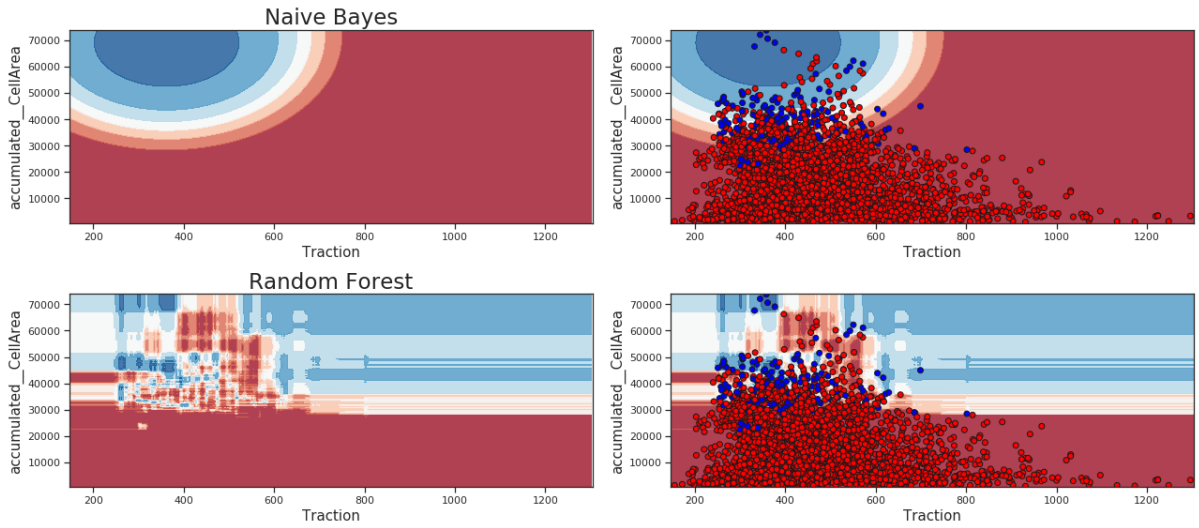


*Figure 11. Gaussian Naive Bayes and Random Forest models trained using the traction force and the cumulative cell area features. In the first column it is shown models behaviour and in the second column our dataset is added to the initial representation.*

Finally, as in the other two cases, Figure 11 shows that the RF Classifier model overfits the training set and Gaussian NB model results are better (seeTable 6).

| | Held-Out Likelihood | F1 Score | Precision | Recall |
|---|---|---|---|---|
| Gaussian Naive Bayes | -449.9 | 0.43 | 0.45 | 0.42 |
| Random Forest Classifier | -441.89 | 0.23 | 0.43 | 0.16 |

*Table 6. Third Comparison between the bests Gaussian NB and RF Classifier models.*

Despite RF model has drastically increased the precision, it has decreased the recall in the same measure. So, I am not going to consider this RF model either.

Comparing the Gaussian NB trained models, using the accumulative cytoplasmic area values and the cumulative Energy 2 values, I get a significantly better recall value and, as a consequence, a better F1 Score, which means that the first model is able to recover a higher number of true samples than the others.

On the other hand, with the other models I get a better precision value, which means that by using these pairs of feature combinations (the pair of tension force and cumulative cell area and the pair of traction force and cumulative cell area) when the models assign a true value to a sample, this prediction is correct in more cases than when this is done by the first model.

Comparing the two last models between each other, the one that uses the tension force and the cumulative cell area features has higher precision, recall and F1 score

values than the one that uses the traction force and the cumulative cell area features, while the second one has a higher held-out likelihood value.

## 4.4. Definition of the Generative Model

Focusing on how NB and RF models behave, the Gaussian Naive Bayes model corresponds to an ellipsoidal-shape function, whose behaviour resembles a propagation wave that is created when throwing a stone in calm water. So In Figure 9 what you can see is that for cytoplasmic areas larger than a certain value and accumulated energies above a certain value the model predicts phase change. A way to try to reproduce this 'step' behaviour i by using a sigmoidal factor. I.e. the probability that a cell divides is a sigmoidal factor of the area minus some reference value for the area. A hyperbolic tangent is an example of easy to handle sigmoidal function. According to the NB model there is a similar behaviour for the accumulated Energy 2, so it makes sense to try to model this as a product of sigmoidal functions. The biological interpretation is that the cells that have both a larger accumulated area and a larger accumulated energy are more likely to divide.

Taking this into account, we performed the first generative model version in which we used the hyperbolic tangent function to try to imitate this sigmoid-shape that the obtained curves have.

The performed likelihood function is:

$$\mathscr{L}=\prod_{t_i;1} p(1|x_1^i,x_2^i)\prod_{t_j;0} p(0|x_1^j,x_2^j)$$

where our parameters are $\mathscr{L}=f(\alpha,X_1,\beta,X_2)$,

and the probability for a sample to be predicted as True ($y_{pred}=1$, the cell is going to change to the S phase within the next five time steps):

$$p(1|x_1^i,x_2^i)=(\frac{1+tgh[\alpha(x_1^i-X_1)]}{2})(\frac{1+tgh[\beta(x_2^i-X_2)]}{2}),$$

so the probability for a sample to be predicted as False ($y_{pred}=0$, the cell will remain in the G1 phase, at least for the next five time steps) is:

$$p(0|x_1^i,x_2^i)=1-p(1|x_1^i,x_2^i)$$

Being $x_1^i, x_2^i$ the values of cumulative cytoplasmic area and cumulative Energy 2 of sample $i$, $X_1$ and $X_2$ can be explained explained as the boundary values of these

features that will identify if each variable has the minimum value to affect the prediction in having a True or a False result. On the other hand, $\alpha$ and $\beta$ are the constants that indicate in which measure each one of the features affects to the final prediction.

## 4.5. Evaluating the Generative Model

Once the generative model was defined I evaluated it in our data set. The first step is to fix the initial values that I will assign to the likelihood function parameters ( $\alpha, X_1, \beta, X_2$). To get the initial $X_1$and $X_2$values (the boundary values for each one of the features that will perform the model) I look for the timestamp in which each cell changes from the G1 to the S phase of the cell cycle and I pick up the instant value of the features that we are using as an input for likelihood function. The value for these constants comes from the average of the values for all cell folds. After that, I calculate the $\alpha$ and $\beta$ values as the inverse value of $X_1$and $X_2$, respectively, divided by 100.

After setting the initial parameter values, I analyse the model with the cell-fold cross validation technique. First, I minimize its error by adjusting these parameters to the training set with the BFGS method (this step has as an output the optimized constants). Then I test our likelihood function model by using the test set and the optimized constants.

I repeated this process for all two-possible features combinations, because in a preliminary analysis the models that used the features we set as the best in previous phases got too bad results. Table 7 shows the features used in generative models that achieved the best results.

| 2 Features Generative Models Results | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Cytoplasmic Area | cumulative Energy | -485.83 | 0.35 | 0.51 | 0.26 |
| cumulative Energy | cumulative Energy 2 roll | -570.40 | 0.41 | 0.52 | 0.34 |
| Energy | cumulative Cell Area | -694.61 | 0.03 | 1.00 | 0.02 |
| cumulative Energy 2 | cumulative Energy 2 roll | -870.61 | 0.36 | 0.38 | 0.35 |

*Table 7. Results obtained by the best Generative models performed.*

For studying the best generative models behaviour, I printed in a figure the best models I obtained (taking into account the held-out likelihood and the F1 score metrics) and how it fits our dataset. Figures 12 and 13 show that the optimized models fit the dataset in a similar way than the Gaussian NB models do, despite this the results achieved are not as good as the ones obtained by the models generated by the NB algorithm.

By using the cumulative cytoplasmic area and the cumulative energy on the likelihood function we obtain the best held-out likelihood value. On the other hand, using the cumulative Energy 2 roll, instead of the cumulative cytoplasmic area, we obtain the best F1 score. This last model, having practically the same precision than the first one, has a highly better recall, which means that is able to recover an 8% more true positive cases with the same precision.

In addition to the metric results obtained, the consistency of the parameter values of the different models reinforce the cumulative Energy as a promising feature for the generation of a model able to predict the G1 to S phase change. In these models explained, the cumulative Energy boundary value ($X_2$ parameter in the first model and $X_1$ in the second one) has the same value (1.27E+07) after model optimization and the constants that specify the importance of the feature in the model ($\beta$ in the first and $\alpha$ in the second), in addition to having the same magnitude order, their values are very close (2.53E-07 and 2.72E-07).

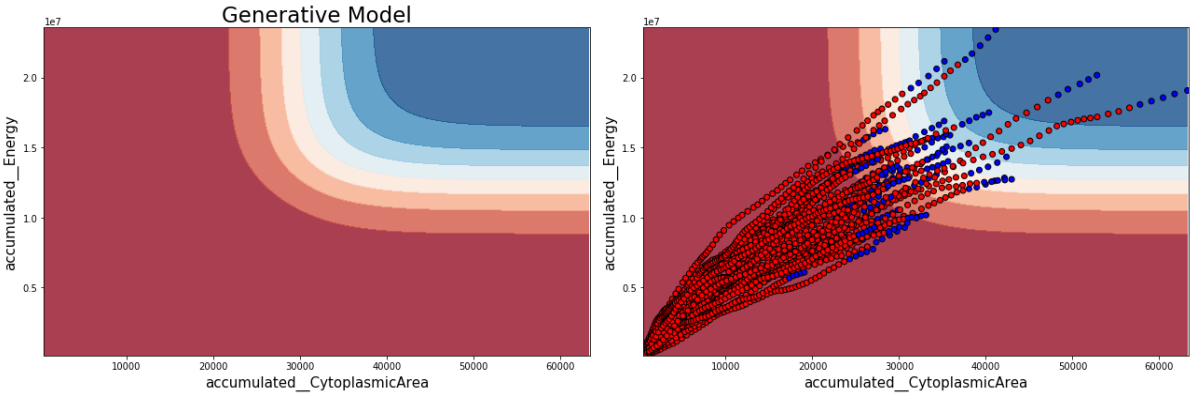


*Figure 12. Generative model trained using the cumulative cytoplasmic area and the cumulative Energy features. In the first column it is shown model behaviour (probability 0 is represented by the colour red and 1 is represented by the colour blue). In the second column our dataset is added (red points mean that the cell is going to remain in the G1 cell cycle phase for the next 5 time stamps, while blue points mean that the cell is going to change from G1 to S phase).*

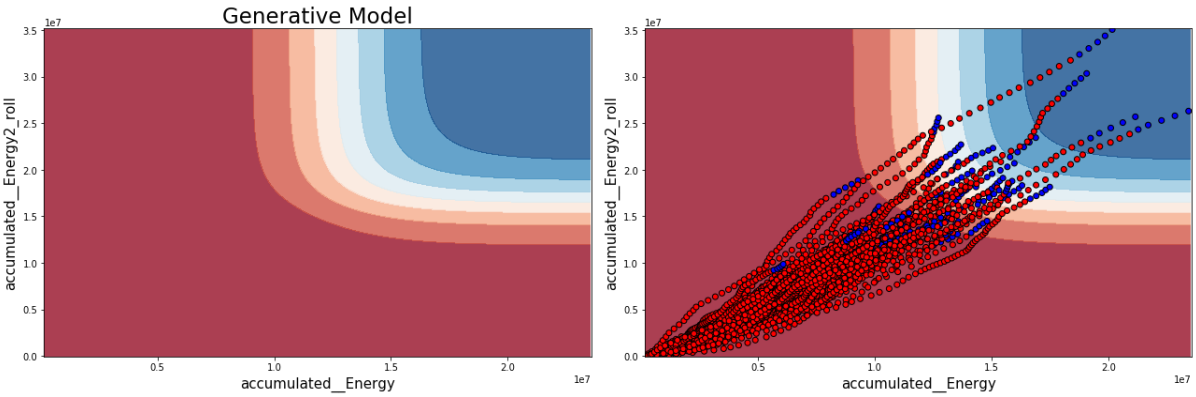


*Figure 13.Generative model trained using the cumulative Energy and the cumulative Energy 2 rolled (the result of the product between the cell area and the traction force) features. In the first column it is shown model behaviour. In the second column our dataset is added.*

## 4.6. Additional Generative Models

The generative models described in the previous section have a satisfactory performance even though they do not quite reproduce the ellipsoidal shape of the Gaussian NB model. Because of this, we designed new models that were in principle mathematically closer to the Gaussian NB function shape, with the expectation that these models would improve prediction metrics.

Specifically, we opted for using a function that had a single sigmoid function instead of two. To account for the balance between biological features as cell division predictors (that is, if one of the two features is high but the other is low, the cell would not divide), the argument of the sigmoid function is then proportional to the product of a power of the two features plus a constant.

The probability for a sample to be predicted as True ($y_{pred}=1$, the cell is going to change to the S phase within the next five time steps) then reads as

$$p(1|x_1^i, x_2^i)=\left(\frac{1+tgh\left[\alpha_0\left(x_1^{i\,\alpha_1}\right)\left(x_2^{i\,\alpha_2}\right)+\alpha_3\right]}{2}\right),$$

where $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ are the parameters of the model. Conversely, the probability for a sample to be predicted as False ($y_{pred}=0$, the cell will remain in the G1 phase, at least for the next five time steps) is

$$p(0|x_1^i, x_2^i) = 1 - p(1|x_1^i, x_2^i)$$

As before we construct the likelihood function for a set of observed data points as:

$$\mathscr{L} = \prod_{t_i;1} p(1|x_1^i, x_2^i) \prod_{t_j;0} p(0|x_1^j, x_2^j)$$

Some of the models obtained using this model have better held-out likelihood value than the best models implemented by the other algorithms, such as the Gaussian NB algorithm or the sigmoid product model from the previous section (see results in Table 8). Even though the model that has the best held-out likelihood value is gained when using the Energy 2 and cumulative cell area features, the model performed by using the cumulative nucleus area, instead of the Energy 2 feature, has better results in the other metrics and its held-out likelihood value is still over the ones obtained by the other models. Unfortunately, the values gained in these other metrics (F1 score, precision and recall) are worse than the ones achieved by the other models.

This function has also generated extreme models which prioritize the precision value, such as the one that uses the traction force and the cumulative energy, or prioritize the recall value, such as the one that uses the tension force and the cumulative cytoplasmic area.

| Best 2nd-Generative Models Results | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Energy 2 | cumulative Cell Area | -436.26 | 0.32 | 0.48 | 0.24 |
| cumulative Cell Area | cumulative Nucleus Area | -440.64 | 0.37 | 0.55 | 0.28 |
| Traction | cumulative Energy | -477.03 | 0.16 | 0.75 | 0.09 |
| Tension | cumulative Cytoplasmic Area | -5512.24 | 0.14 | 0.07 | 1.00 |

*Table 8. Results obtained by the best 2nd-Generative models performed.*

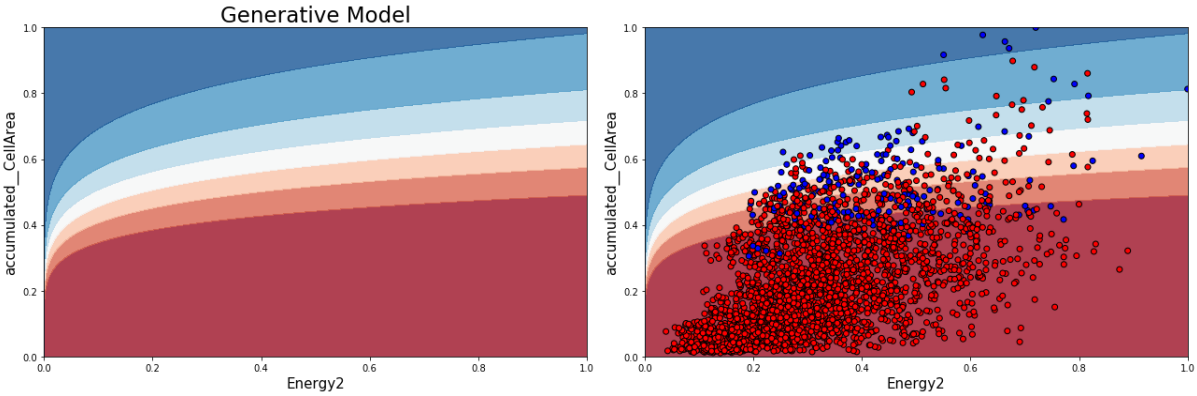


*Figure 14. Generative model trained using the Energy 2 and the cumulative cell area features. In the first column it is shown model behaviour (the gradient colour indicates the prediction value, where probability 0 is represented by the colour red and 1 is represented by the colour blue). In the second column our dataset is added (red points mean that the cell is going to remain in the G1 cell cycle phase for the next 5 time stamps, while blue points mean that the cell is going to change from G1 to S phase in the next 5 time stamps).*

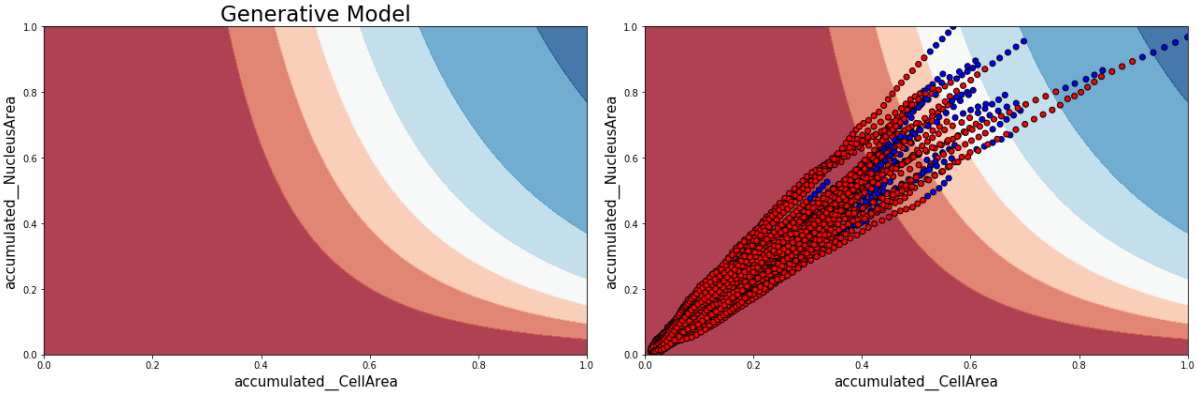


*Figure 15. Generative model trained using the cumulative cell area and the cumulative nucleus area features. In the first column it is shown model behaviour. In the second column our dataset is added.*

Because the product of sigmoid functions seems to give better results that a single sigmoid, we also tried a model with a product of sigmoid functions that have as arguments the powers of the biological features. In this case,

$$\mathscr{L} = \prod_{t_i;1} p\left(1|x_1^i, x_2^i\right) \prod_{t_j;0} p\left(0|x_1^j, x_2^j\right)$$

the probability for a sample to be predicted as True is

$$p\left(1|x_1^i, x_2^i\right) = \left(\frac{1+tgh\left[\alpha_0\left(x_1^{i\,\alpha_1}\right)+\alpha_2\right]}{2} \times \frac{1+tgh\left[\alpha_3\left(x_2^{i\,\alpha_4}\right)+\alpha_5\right]}{2}\right),$$

where the model parameters have now increased from 4 to 6 $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5$.

so the probability for a sample to be predicted as False is:

$$p(0|x_1^i, x_2^i) = 1 - p(1|x_1^i, x_2^i)$$

The performance of the model with all possible two-feature combinations showed two drastic cases. The first one, in which the model that uses the cumulative nucleus area and the cumulative cytoplasmic area achieves the maximum precision at the cost of having an awful recall, and another one, in which the models achieve the maximum recall at the cost of having an awful precision. (Table 9 only shows one of all the two-pair features models that achieve this recall value of 1.00, see the rest in Tables 21 and 22 Appendix H).

| 3rd-Generative Models Results | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -504.64 | 0.05 | 1.00 | 0.03 |
| Cell Area | Tension | -5224.73 | 0.14 | 0.07 | 1.00 |

Table 9. Results obtained by the best 3rd-Generative models performed.

The behaviour of these models' performance suggests that only one of the features are taken into account when doing the predictions. The first generative model only takes into account the cumulative nucleus area (Figure 16) and the second the cell area (Figure 17).

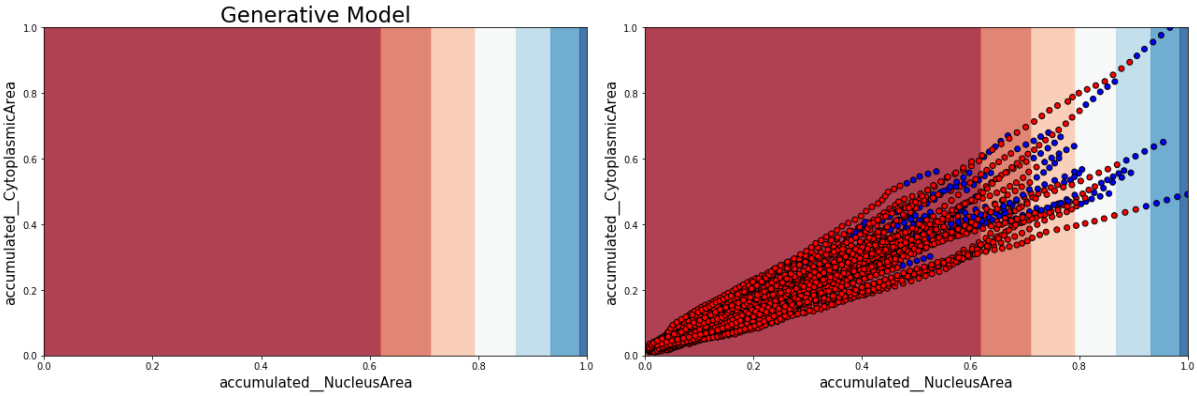


*Figure 16. Generative model trained using the cumulative cytoplasmic area and the cumulative nucleus area features. In the first column it is shown model behaviour (the gradient colour indicates the prediction value, where probability 0 is represented by the colour red and 1 is represented by the colour blue). In the second column our dataset is added (red points mean that the cell is going to remain in the G1 cell cycle phase for the next 5 time stamps, while blue points mean that the cell is going to change from G1 to S phase in the next 5 time stamps).*

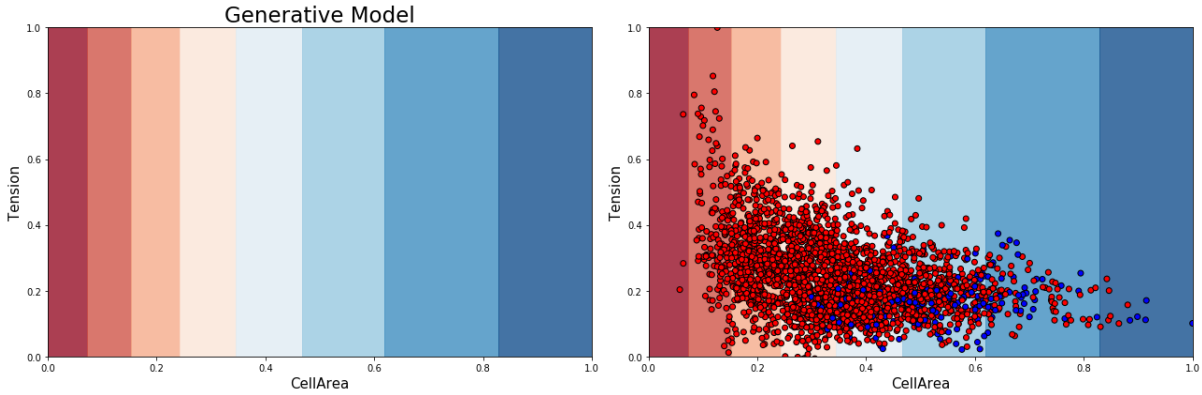


*Figure 17. Generative model trained using the cell area and the tension force features. In the first column it is shown model behaviour. In the second column our dataset is added.*

## 4.7. Future Work

The next phase of the project consists in trying to add an extra feature to our performed models to try to increase its prediction results. The methodology to do that should be the following:

1. Study which is the feature than improves our best Gaussian NB performed models by adding it into the training phase.
2. Study the performed model behaviour. As now we will have a 3D model, the way to do this step is generating many figures where in each one of them are represented the probabilities by modifying the values of two features and letting the third one as a constant value.
3. Extrapolate the results observed to the likelihood functions by adding the third variable in the proper way.
4. Perform, with the new likelihood functions, and evaluate these new models.

# 5. Conclusions

The possibility of having a model able to predict when a cell is going to move from one cell cycle phase to another and if it has an accelerated or lowered cell cycle can be a useful tool in the treatment of tissue proliferation diseases such as cancer.

Even though results in this project show that, for achieving this purpose, the best machine learning tool is the Gaussian Naive Bayes algorithm, generative models while they are not as predicitve in the current version as Gaussian NB are promising because of their simplicity and easy biological formulation and interpretation. As I mentioned in the last section, a way to achieve this goal is by properly adding another variable to the actual likelihood functions.

Another interesting thing that could improve our models performance is the availability of more data. When using machine learning methods, having a big amount of data is important because the algorithm examines more cases and can obtain better models. For example, with more data we could in principle solve the overfitting problem in the Random Forest Classifier models. In the same scope, another thing that could affect the development of better models is the fact that our dataset has a huge amount of False samples (cell will remain in the G1 phase) respect to the number of True samples (cell is going to change from the G1 phase to the S phase). An option to solve this issue is predicting the transition in the next 10 time steps instead of 5, which would duplicate the number of True samples in the dataset.

Despite the current limitations of my analysis, results show that the best predictors are cumulative areas, the instantaneous traction and tension forces and the cumulative energies, which is consistent the results achieved by Uroz, M. *et al.*. In this manuscript, the authors proposed the accumulation of mechanical energy via mechanical stress (our cumulative Energy feature) as an explanation for the duration of the G1 phase. My results on the models based on cumulative cytoplasmic area and cumulative Energy features, show that these are the best features for prediction of change of phase in cell division as well.

Finally, based on all the experience I have gained during the project, in three-feature

models, I would combine two cumulative features with an instantaneous one: a first one that contains information about the area; second one that carries information the force features; and a last one that is an energy feature (presumably the one that is not calculated by using the force selected as the second variable).

# 6. Self evaluation

The personal valoration about the work I have realized in the SEES Lab is, without any doubt, positive. It has allowed me to put in common in the same project all the knowledge acquired from both degrees (Computer Science and Biotechnology) during the last 5 years. On the one hand, the experience has helped me on the understanding of the data with which I worked and explaining the results achieved as a consequence of processes carried out during the cell cycle. On the other hand, I had a better comprehension of the algorithms that implement the machine learning methods used (because I studied some of them during the degree), what helped me interpreting the results given. In addition, as I had worked before with the programming language used to carry out all the experiments and I knew most of the data structures used, I was able to start working without previous programming lessons what agilitzed the work.

During the project, there have been periods in which everything was booming and others in which I have been discouraged because nothing was going as expected. This experience has shown me what the research world is and I have fallen in love with it, this what I want to do in the future. Every day was a challenge, for example: check the results of the experiments that I started the previous day, which could had been executing during the night or during the weekend, discover they are useless and try to find another path to see the light at the end of the tunnel.

Finally, I just wanted to say a big thank you Marta Sales and Roger Guimerà for all the assistance provided during the project as well as for the trust they have put on me. Also, I would like to thank Sergio Cobo, PhD researcher in the SEES Lab, for helping me in those moments in which I was discouraged with the work.

# 7. Bibliography


Alberts, B. *et al*. (2002). Molecular Biology of the Cell. 4th edition, *Annals of Botany* **91**(3), 401. https://doi.org/10.1093/aob/mcg023

Aragona, M. *et al.* 2013. A mechanical checkpoint controls multicellular growth through YAP/TAZ regulation by actin-processing factors. *Cell* **154**(5), 1047–1059. https://doi.org/10.1016/j.cell.2013.07.042

Bartek, J. and Lukas, J. (2001). Pathways governing G1/S transition and their response to DNA damage, *FEBS Letters* **490** (3), 117–22. https://doi.org/10.1016/S0014-5793(01)02114-7

Benham-Pyle, B. *et al.* 2015. Cell adhesion. Mechanical strain induces E-cadherin-dependent Yap1 and β-catenin activation to drive cell cycle entry. *Science* **348**(6238), 1024-1027. https://doi.org/10.1126/science.aaa4559

Bertoli, C. *et al*. (2013). Control of cell cycle transcription during G1 and S phases, *Nature Reviews Molecular Cell Biology* **14**(8), 518–28. https://doi.org/10.1038/nrm3629

Chen, C. *et al.* (1997). Geometric control of cell life and death, *Science* **276**(5317), 1425-8. https://doi.org/10.1126/science.276.5317.1425

Chinchor, N. (1992). MUC-4 Evaluation Metrics, *in Proc. of the Fourth Message Understanding Conference*, pp. 22–29. https://www.aclweb.org/anthology/M92-1002

Folkman, J. and Moscona, A. (1978). Role of cell shape in growth control, *Nature* **273**(5661), 345-349. https://www.nature.com/articles/273345a0

Gudipaty, S. *et al.* 2017. Mechanical stretch triggers rapid epithelial cell division through Piezo1. *Nature* **543**(7643), 118–121. https://doi.org/10.1038/nature21407

Huang, S. *et al*. (1998). Control of Cyclin D1, p27 Kip1 , and Cell Cycle Progression in Human Capillary Endothelial Cells by Cell Shape and Cytoskeletal Tension, *Molecular Biology of the Cell* **9**(11), 3179-3193. https://doi.org/10.1091/mbc.9.11.3179

Joanneum, F. (2005-2006). Cross-Validation Explained, *Institute for Genomics and Bioinformatics.* http://genome.tugraz.at/proclassify/help/pages/XV.html

Lancaster, O. *et al*. 2013. Mitotic rounding alters cell geometry to ensure efficient bipolar spindle formation. *Dev. Cell* **25**(3), 270-83. https://doi.org/10.1016/j.devcel.2013.03.014

LeGoff, L. and Lecuit, T. 2015. Mechanical forces and growth in animal tissues. *Cold Spring Harb. Perspect. Biol.* **8**(3), a019232. https://doi.org/10.1101/cshperspect.a019232

Lloyd, A. (2013). The regulation of cell size, *Cell* **154**(6), 1194-205. https://doi.org/10.1016/j.cell.2013.08.053

Mih, J. *et al*. 2012. Matrix stiffness reverses the effect of actomyosin tension on cell proliferation. *J. Cell Sci.* **125**(Pt 24), 5974–5983. https://doi.org/10.1242/jcs.108886

Pinheiro, D. *et al.* 2017. Transmission of cytokinesis forces via E-cadherin dilution and actomyosin flows. *Nature* **545**(7652), 103–107. https://doi.org/10.1038/nature22041

Roca-Cusachs, P. *et al*. (2008). Micropatterning of single endothelial cell shape reveals a tight coupling between nuclear volume in G1 and proliferation, *Biophysical Journal* **94**(12), 4984–4995. https://doi.org/10.1529/biophysj.107.116863

Sasaki, Y. (2007). The truth of the F-measure. https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure

Saxton, R. and Sabatini, D. (2017). mTOR Signaling in Growth, Metabolism, and

Disease, *Cell* **168**(6), 960-976. https://doi.org/10.1016/j.cell.2017.02.004

Son, S. *et al*. (2012). Direct observation of mammalian cell growth and size regulation, *Nature Methods* **9**(9), 910–912. https://doi.org/10.1038/nmeth.2133

Streichan, S. *et al.* 2014. Spatial constraints control cell proliferation in tissues. *Proc. Natl Acad. Sci.* **111**(15) 5586-5591. https://doi.org/10.1073/pnas.1323016111

Uroz, M. *et al*. (2018). Regulation of cell cycle progression by cell–cell and cell–matrix forces, *Nature Cell Biology* **20**(6), 646–654. https://doi.org/10.1038/s41556-018-0107-2

Utgoff, P. E. (1989). Incremental Induction of Decision Trees, *Machine learning* **4**(2), 161-186. https://doi.org/10.1023/A:1022699900025

Vianay, B. *et al.* 2018. Variation in traction forces during cell cycle progression. *Biol. Cell* **110**(4):91-96. https://doi.org/10.1111/boc.201800006

Wallach, H. *et al.* (2009). Evaluation Methods for Topic Models, *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, 1105-1112. https://doi.org/10.1145/1553374.1553515

Watt, F. *et al.* (1988). Cell shape controls terminal differentiation of human epidermal keratinocytes, *Proc. Natl Acad. Sci.* **85**(15), 5576-80. https://doi.org/10.1073/pnas.85.15.5576

Yu, FX., *et al.* (2015). Hippo Pathway in Organ Size Control, Tissue Homeostasis, and Cancer, *Cell 163*(4), 811-828. https://doi.org/10.1016/j.cell.2015.10.044

Zhang, H. (2004). The Optimality of Naive Bayes. *Proceedings of 17th International Florida Artificial Intelligence Research Society Conference, Menlo Park*, 562-567. https://www.researchgate.net/publication/221439320_The_Optimality_of_Naive_Bayes

# Appendix A: Data

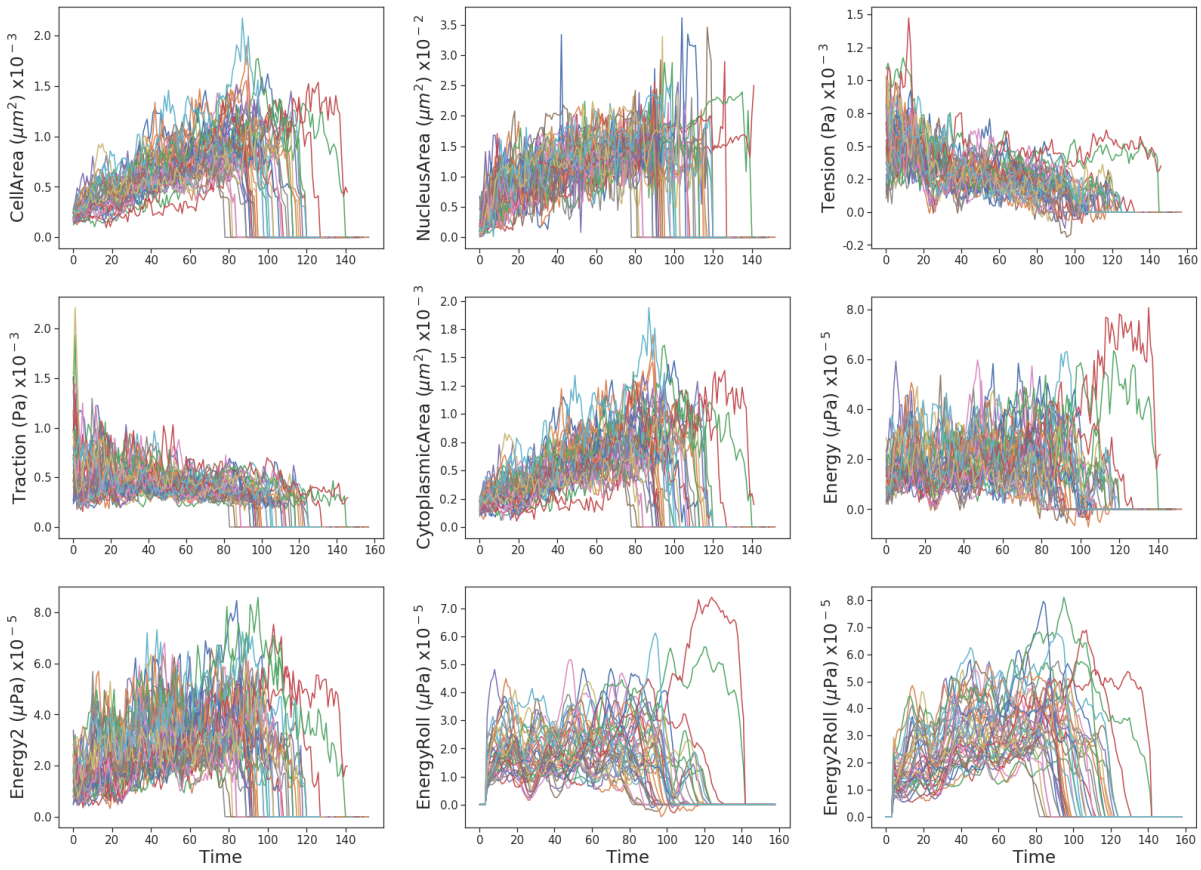


*Figure 18. Representation over the time of all the calculated features of the cells in the dataset. There is a moment from which the features get a 0 value, this is because, from that timestamp, there are no further measures. Energy corresponds to the product between the cell size and the cell-cell tension. Energy 2 corresponds to the product between the cell size and the cell-ECM traction. The "Roll" Energies are calculated in the same way as the normal Energy but for each value I use a rolling window with the 5 last values.*

# Appendix B: Feature correlations

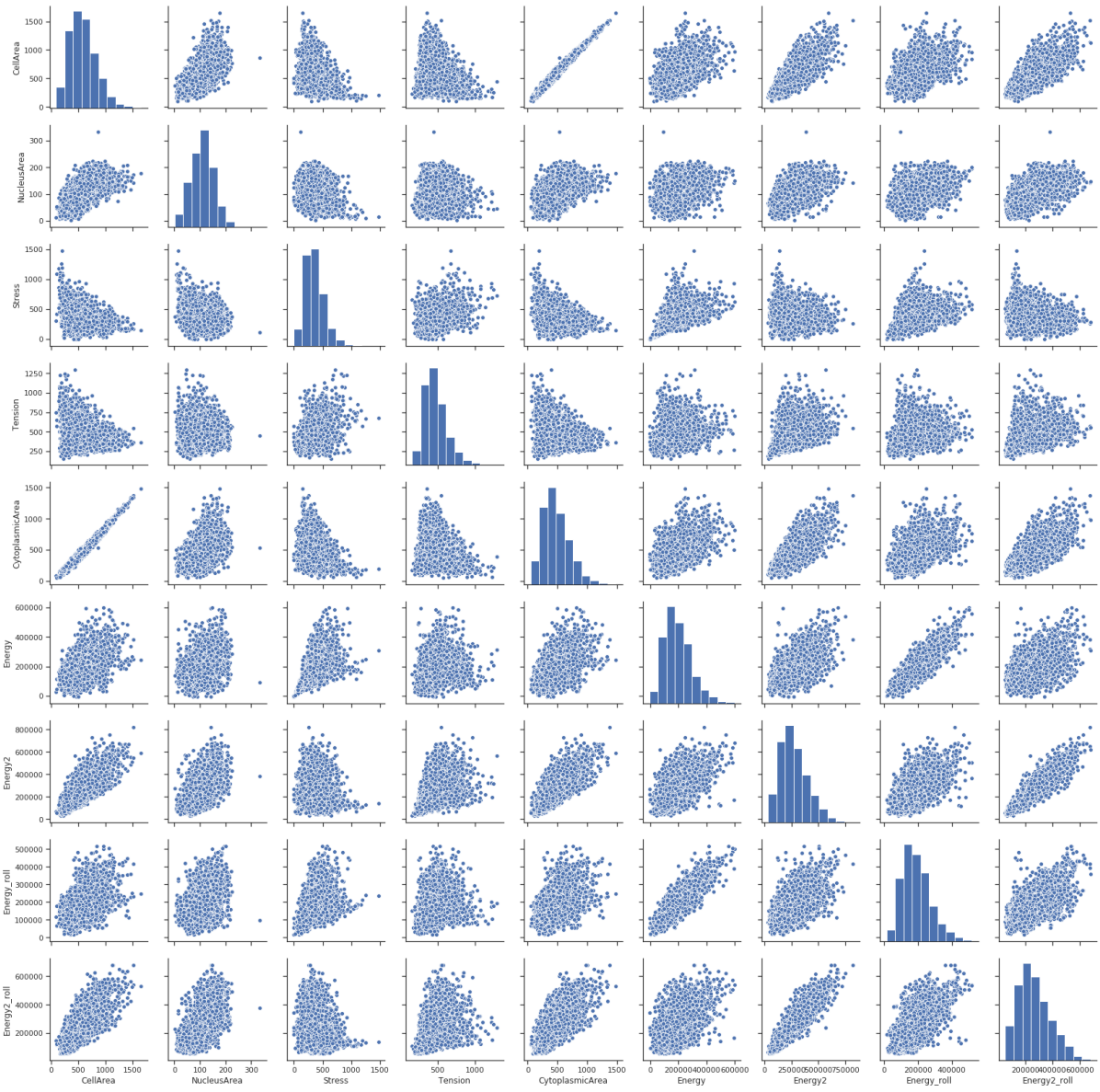


*Figure 19. Feature correlations. The features are from left to right and from top to bottom: Cell Area, Nucleus Area, Tension force, Traction force, Cytoplasmic Area, Energy (product between the cell size and the cell-cell tension), Energy 2 (product between the cell size and the cell-ECM traction), Energy roll (calculated as the Energy value but using a rolling window with the 5 last values instead of the instantaneous value) and Energy 2 roll (calculated as the Energy 2 value but using a rolling window with the 5 last values instead of the instantaneous value).*

# Appendix C: Logistic Regression models results

| Logistic Regression Models Results sorted by the Held-Out likelihood value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Traction | cumulative Nucleus Area | -452.47 | 0.38 | 0.65 | 0.27 |
| Nucleus Area | cumulative Cytoplasmic Area | -459.74 | 0.26 | 0.53 | 0.18 |
| Nucleus Area | cumulative Nucleus Area | -465.06 | 0.32 | 0.56 | 0.23 |
| Cell Area | cumulative Nucleus Area | -465.58 | 0.29 | 0.55 | 0.20 |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -465.97 | 0.33 | 0.56 | 0.23 |
| cumulative Cell Area | cumulative Nucleus Area | -466.03 | 0.27 | 0.56 | 0.18 |
| Cytoplasmic Area | cumulative Nucleus Area | -466.14 | 0.28 | 0.54 | 0.19 |
| Tension | cumulative Nucleus Area | -468.46 | 0.28 | 0.59 | 0.19 |
| Cell Area | cumulative Cell Area | -470.29 | 0.27 | 0.52 | 0.19 |
| Cell Area | cumulative Cytoplasmic Area | -482.40 | 0.29 | 0.53 | 0.20 |

*Table 11. Best 10 LR models results sorted by the held-out likelihood value.*

| Logistic Regression Models Results sorted by the F1 score value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Traction | cumulative Nucleus Area | -452.47 | 0.38 | 0.65 | 0.27 |
| cumulative Energy 2 | cumulative Energy 2 roll | -598.11 | 0.33 | 0.45 | 0.26 |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -465.97 | 0.33 | 0.56 | 0.23 |
| Nucleus Area | cumulative Nucleus Area | -465.06 | 0.32 | 0.56 | 0.23 |
| Energy 2 roll | cumulative Energy 2 roll | -706.89 | 0.32 | 0.40 | 0.27 |
| Energy 2 roll | cumulative Cytoplasmic Area | -767.10 | 0.31 | 0.38 | 0.27 |
| Traction | cumulative Cell Area | -528.00 | 0.31 | 0.48 | 0.23 |
| Energy 2 roll | cumulative Energy 2 | -718.01 | 0.30 | 0.39 | 0.25 |
| Energy 2 roll | cumulative Cell Area | -771.20 | 0.30 | 0.36 | 0.26 |
| Cell Area | cumulative Nucleus Area | -465.58 | 0.29 | 0.55 | 0.20 |

*Table 12. Best 10 LR models results sorted by the F1 Score value.*

# Appendix D: Random Forest Classifier models results

| 2 Features Random Forest Models Results sorted by the Held-Out likelihood value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Traction | cumulative Cytoplasmic Area | -423.85 | 0.10 | 0.19 | 0.07 |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -424.43 | 0.20 | 0.41 | 0.14 |
| cumulative Tension | cumulative Cytoplasmic Area | -429.75 | 0.33 | 0.60 | 0.23 |
| Traction | cumulative Cytoplasmic Area | -431.18 | 0.16 | 0.35 | 0.11 |
| cumulative Cell Area | cumulative Tension | -431.93 | 0.36 | 0.62 | 0.25 |
| CellArea | cumulative Cytoplasmic Area | -433.36 | 0.18 | 0.35 | 0.13 |
| Cytoplasmic Area | cumulative Cytoplasmic Area | -435.87 | 0.11 | 0.23 | 0.08 |
| cumulative Cytoplasmic Area | cumulative Energy roll | -437.46 | 0.21 | 0.40 | 0.15 |
| CellArea | cumulative Cell Area | -437.84 | 0.17 | 0.34 | 0.11 |
| cumulative Cell Area | cumulative Energy 2 roll | -438.48 | 0.28 | 0.42 | 0.21 |

*Table 13. Best 10 RF Classifiers models results sorted by the held-out likelihood value.*

| 2 FeaturesRandom Forest Models Results sorted by the F1 score value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Cell Area | cumulative Tension | -431.93 | 0.36 | 0.62 | 0.25 |
| cumulative Tension | cumulative Cytoplasmic Area | -429.75 | 0.33 | 0.60 | 0.23 |
| Energy 2 roll | cumulative Energy | -inf | 0.28 | 0.55 | 0.19 |
| cumulative Cell Area | cumulative Energy 2 roll | -438.48 | 0.28 | 0.42 | 0.21 |
| Energy 2 roll | cumulative Energy roll | -inf | 0.27 | 0.53 | 0.19 |
| Energy 2 roll | cumulative Cell Area | -inf | 0.27 | 0.50 | 0.19 |
| cumulative Cell Area | cumulative Energy 2 | -inf | 0.27 | 0.43 | 0.20 |
| cumulative Nucleus Area | cumulative Traction | -475.72 | 0.25 | 0.54 | 0.17 |
| Cytoplasmic Area | cumulative Nucleus Area | -484.63 | 0.24 | 0.54 | 0.16 |
| Traction | cumulative Cell Area | -441.89 | 0.23 | 0.43 | 0.16 |

*Table 14. Best 10 RF Classifiers models results sorted by the F1 Score value.*

# Appendix E: Gaussian Naive Bayes Models result

| Gaussian Naive Bayes Models Results sorted by the Held-Out likelihood value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Traction | cumulative Cell Area | -449.62 | 0.43 | 0.45 | 0.42 |
| Energy | cumulative Cell Area | -460.09 | 0.37 | 0.44 | 0.33 |
| Energy roll | cumulative Cell Area | -462.88 | 0.35 | 0.40 | 0.32 |
| Tension | cumulative Cell Area | -463.94 | 0.46 | 0.46 | 0.46 |
| Traction | cumulative Cytoplasmic Area | -464.73 | 0.41 | 0.44 | 0.39 |
| Traction | cumulative Nucleus Area | -474.64 | 0.42 | 0.44 | 0.41 |
| Energy | cumulative Cytoplasmic Area | -475.90 | 0.37 | 0.44 | 0.32 |
| Nucleus Area | cumulative Cell Area | -476.73 | 0.46 | 0.45 | 0.47 |
| Tension | cumulative Cytoplasmic Area | -477.85 | 0.42 | 0.43 | 0.41 |
| Energy roll | cumulative Cytoplasmic Area | -478.67 | 0.35 | 0.40 | 0.31 |

*Table 15. Best 10 Gaussian NB models results sorted by the held-out likelihood value.*

| Gaussian Naive Bayes Models Results sorted by the F1 Score value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -653.92 | 0.50 | 0.40 | 0.67 |
| cumulative Cell Area | cumulative Cytoplasmic Area | -688.64 | 0.49 | 0.40 | 0.65 |
| cumulative Cell Area | cumulative Nucleus Area | -671.53 | 0.49 | 0.39 | 0.68 |
| cumulative Cytoplasmic Area | cumulative Energy 2 | -706.27 | 0.48 | 0.39 | 0.60 |
| cumulative Cytoplasmic Area | cumulative Energy | -660.89 | 0.48 | 0.39 | 0.61 |
| cumulative Cytoplasmic Area | cumulative Energy 2 roll | -710.90 | 0.47 | 0.39 | 0.60 |
| cumulative Cytoplasmic Area | cumulative Energy roll | -666.26 | 0.47 | 0.38 | 0.60 |
| cumulative Cell Area | cumulative Energy | -659.79 | 0.47 | 0.38 | 0.61 |
| cumulative Cell Area | cumulative Energy 2 | -703.16 | 0.46 | 0.38 | 0.60 |
| cumulative Cell Area | cumulative Energy 2 roll | -708.51 | 0.46 | 0.38 | 0.60 |

*Table 16. Best 10 Gaussian NB models results sorted by the F1 Score value.*

# Appendix F: Generative models results

| Generative Models Results sorted by the Held-Out likelihood value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Cytoplasmic Area | cumulative Energy | -485.83 | 0.35 | 0.51 | 0.26 |
| cumulative Cell Area | cumulative Energy | -487.07 | 0.36 | 0.52 | 0.27 |
| cumulative Cytoplasmic Area | cumulative Energy roll | -487.81 | 0.38 | 0.54 | 0.29 |
| cumulative Nucleus Area | cumulative Energy 2 roll | -488.80 | 0.38 | 0.52 | 0.30 |
| cumulative Cell Area | cumulative Energy roll | -489.22 | 0.38 | 0.53 | 0.29 |
| cumulative Tension | cumulative Energy 2 roll | -492.37 | 0.20 | 0.53 | 0.12 |
| cumulative Cytoplasmic Area | cumulative Energy 2 roll | -494.98 | 0.34 | 0.53 | 0.25 |
| cumulative Cytoplasmic Area | cumulative Energy 2 | -498.15 | 0.32 | 0.48 | 0.24 |
| cumulative Nucleus Area | cumulative Energy 2 | -499.08 | 0.32 | 0.46 | 0.25 |
| cumulative Cell Area | cumulative Energy 2 roll | -508.98 | 0.31 | 0.47 | 0.24 |

*Table 17. Best 10 Generative models results sorted by the held-out likelihood value.*

| Generative Models Results sorted by the F1 Score value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Energy | cumulative Energy 2 roll | -570.40 | 0.41 | 0.52 | 0.34 |
| cumulative Energy roll | cumulative Energy 2 roll | -577.71 | 0.40 | 0.49 | 0.34 |
| cumulative Energy 2 | cumulative Energy roll | -572.30 | 0.39 | 0.49 | 0.33 |
| cumulative Nucleus Area | cumulative Energy 2 roll | -488.80 | 0.38 | 0.52 | 0.30 |
| cumulative Cytoplasmic Area | cumulative Energy roll | -487.81 | 0.38 | 0.54 | 0.29 |
| cumulative Cell Area | cumulative Energy roll | -489.22 | 0.38 | 0.53 | 0.29 |
| cumulative Energy 2 | cumulative Energy 2 roll | -870.61 | 0.36 | 0.38 | 0.35 |
| cumulative Energy | cumulative Energy 2 | -645.69 | 0.36 | 0.46 | 0.30 |
| cumulative Cell Area | cumulative Energy | -487.07 | 0.36 | 0.52 | 0.27 |
| cumulative Cytoplasmic Area | cumulative Energy | -485.83 | 0.35 | 0.51 | 0.26 |

*Table 18. Best 10 Generative models results sorted by the F1 Score value.*

# Appendix G: 2nd-version Generative models results

| 2nd-version Generative Models Results sorted by the Held-Out likelihood value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Energy 2 | cumulative Cell Area | -436.26 | 0.32 | 0.48 | 0.24 |
| Energy 2 roll | cumulative Cytoplasmic Area | -438.19 | 0.33 | 0.50 | 0.25 |
| cumulative Cell Area | cumulative Energy 2 | -438.95 | 0.33 | 0.49 | 0.25 |
| Cytoplasmic Area | cumulative Cell Area | -439.01 | 0.33 | 0.51 | 0.24 |
| Cell Area | cumulative Cell Area | -439.06 | 0.32 | 0.51 | 0.24 |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -439.71 | 0.37 | 0.55 | 0.28 |
| cumulative Cell Area | cumulative Energy 2 roll | -439.75 | 0.33 | 0.49 | 0.25 |
| cumulative Cell Area | cumulative Nucleus Area | -440.64 | 0.37 | 0.55 | 0.28 |
| Nucleus Area | cumulative Cell Area | -441.66 | 0.28 | 0.46 | 0.21 |
| cumulative Cell Area | cumulative Energy | -442.46 | 0.29 | 0.45 | 0.21 |

*Table 19. Best 10 2nd-version Generative models results sorted by the held-out likelihood value.*

| 2nd-version Generative Models Results sorted by the F1 Score value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Cell Area | cumulative Nucleus Area | -440.64 | 0.37 | 0.55 | 0.28 |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -439.71 | 0.37 | 0.55 | 0.28 |
| Traction | cumulative Nucleus Area | -443.45 | 0.35 | 0.61 | 0.25 |
| Energy 2 roll | cumulative Cytoplasmic Area | -438.19 | 0.33 | 0.50 | 0.25 |
| cumulative Cell Area | cumulative Energy 2 roll | -439.75 | 0.33 | 0.49 | 0.25 |
| Cytoplasmic Area | cumulative Cell Area | -439.01 | 0.33 | 0.51 | 0.24 |
| cumulative Cell Area | cumulative Energy 2 | -438.95 | 0.33 | 0.49 | 0.25 |
| cumulative Cell Area | cumulative Tension | -448.67 | 0.33 | 0.49 | 0.25 |
| Cell Area | cumulative Cell Area | -439.06 | 0.32 | 0.51 | 0.24 |
| Energy 2 | cumulative Cell Area | -436.26 | 0.32 | 0.48 | 0.24 |

*Table 20. Best 10 2nd-version Generative models results sorted by the F1 Score value.*

# Appendix H: 3rd-version Generative models results

| 4rt-version Generative Models Results sorted by the Held-Out likelihood value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| cumulative Nucleus Area | cumulative Cytoplasmic Area | -504.64 | 0.05 | 1.00 | 0.03 |
| cumulative Nucleus Area | cumulative Energy 2 | -523.36 | 0.00 | 0.00 | 0.00 |
| cumulative Nucleus Area | cumulative Energy 2 roll | -534.35 | 0.00 | 0.00 | 0.00 |
| cumulative Nucleus Area | cumulative Energy roll | -556.74 | 0.00 | 0.00 | 0.00 |
| cumulative Nucleus Area | cumulative Energy | -556.83 | 0.00 | 0.00 | 0.00 |
| cumulative Cell Area | cumulative Energy 2 roll | -575.64 | 0.00 | 0.00 | 0.00 |
| cumulative Cell Area | cumulative Energy 2 | -580.23 | 0.00 | 0.00 | 0.00 |
| cumulative Nucleus Area | cumulative Traction | -594.58 | 0.00 | 0.00 | 0.00 |
| cumulative Nucleus Area | cumulative Tension | -602.68 | 0.00 | 0.00 | 0.00 |
| cumulative Cell Area | cumulative Traction | -623.27 | 0.00 | 0.00 | 0.00 |

*Table 21: Best 10 3rd-version Generative models results sorted by the held-out likelihood value.*

| 3rd-version Generative Models Results sorted by the F1 Score value | | | | | |
|---|---|---|---|---|---|
| Features | | Held-out likelihood | F1 Score | Precision | Recall |
| Cell Area | Tension | -5224.73 | 0.14 | 0.07 | 1.00 |
| Cell Area | Energy | -5224.73 | 0.14 | 0.07 | 1.00 |
| Nucleus Area | Tension | -5128.74 | 0.14 | 0.07 | 1.00 |
| Nucleus Area | Energy | -5128.74 | 0.14 | 0.07 | 1.00 |
| Tension | Traction | -4834.21 | 0.14 | 0.07 | 1.00 |
| Tension | Cytoplasmic Area | -4834.21 | 0.14 | 0.07 | 1.00 |
| Tension | Energy | -4834.21 | 0.14 | 0.07 | 1.00 |
| Tension | Energy 2 | -4834.21 | 0.14 | 0.07 | 1.00 |
| Tension | Energy roll | -4834.21 | 0.14 | 0.07 | 1.00 |
| Tension | Energy 2 roll | -4834.21 | 0.14 | 0.07 | 1.00 |

*Table 22: Best 10 3rd-version Generative models results sorted by the F1 Score value.*