Deep Learning techniques for cell stage classification

Rafael Freitas
rafaeljcfreitas@ist.utl.pt

Instituto Superior Técnico, Lisboa, Portugal

September 2020

Abstract

The various phases and mechanisms that happen sequentially during a life of a cell form the cell cycle and the correct progression along this cycle is essential for the maintenance of life. For a typical eukaryotes cell this cycle can be separated into 2 main phases: interphase, during which the cell is growing, and mitosis, where the cell separates into two daughter cells. The interphase can be further divided into 3 main phases, G1, where the cell is growing, the S phase, where the DNA is replicated and the G2 phase during which the cell continues growing in preparation for mitosis. Due to the importance of the cellular cycle, the correct staging of a cell (i.e. correct classification of its current phase) is of the utmost importance for biological and pharmacological research. However, current methods for cell staging have traditionally relied on fluorescent microscopy and the analysis of cell population and present some drawbacks such as the need for specific biological markers or the destruction of the cell culture to determine its stage. As such, in this project, a new method that relies on the use of DAPI stained cell cultures (one of the most common cell imaging techniques) and deep learning techniques is proposed. By using deep learning algorithms, this method is capable of staging single data points without relying on cell population analysis and will not require specific biological markers, resulting in fairly simpler process to achieve cell staging.

Keywords: Deep learning, DAPI staining,Fluorescent Microscopy, Cell Cycle, Convolutional Neural Network

1. Introduction

The cell cycle is one of the essential mechanisms that allow the maintenance of life. The correct progression along the cell cycle makes cell growth and reproduction possible by achieving its separation into two daughter cells with equal genome.

Across the cell cycle genome stability is maintained by regulation that establishes not only the correct genome duplication, but also the appropriate chromosomal distribution to each daughter cell. Failure to regulate may cause genome instability which has been linked to abnormal cellular behavior, such as unscheduled proliferation, and to diseases, one such example being cancer.

As such, the study of cell cycle progression is of the utmost importance as is the ability to determine at which stage a cell is, i.e the ability to stage a cell. Most cell staging methods rely on population-based analysis making them incompatible with single-cell staging.

Recently, approaches to this problem developed methods that can accurately stage a single cell or track its progression throughout the cell cycle. Even tough useful, most of these methods require the use of specific cell markers or the use of stage specific reporters meaning they are still incapable of tracking a cell's progression through the cell cycle or are incompatible with modern high resolution biological techniques. [1, 2, 3]

As such, the need for an inclusive method capable of determining any cell cycle stage that is compatible with modern biological techniques is still present.

This project aims to evaluate cells through 40,60-diamidino-2-phenylinodole (DAPI) staining with fluorescence microscopy images and classify them by identifying the corresponding cell cycle phase, considering intracellular features. While some approaches have been developed they relied on the clustering classification and as such could only classify a sufficiently large set of data points. By utilizing deep learning methods, this project hopes to develop a method that can successfully identify single data points after careful training is performed with the original dataset.

2. Background

The following section will provide an overview of essential concepts from both biology and machine learning required to understand the work proposed.

2.1.1 The Cell

Cells are the basic unit of life and are considered the building blocks of all living organisms. Much like these organisms, cells have evolved and adapted to various environments and functional roles. Nonetheless, all cells rely on, basically, the same structures to perform the set of tasks that is essential for their own survival.

One of these common structures that all cells possess are **nucleic acids**, the molecules responsible for containing and helping to express a cell's genetic code). These acids can be classified in two major classes **deoxyribonucleic acid** (DNA), that contains the essential information for the creation and maintenance of the cell and **ribonucleic acid** (RNA), that possesses several roles associated with the expression of genetic material. The DNA is packaged differently in different cells which resulted in a form of cell classification, if the DNA presents itself separated from the cytoplasm by a membrane, than the cell is **eukaryote**. In contrast, if the DNA is in contact with the cytoplasm, without any barrier, than the cell is **prokaryote**. Eukaryotic cells form all animals and plants. As such, since the data used on this thesis refers to human cells, these will be the only ones of interest. Consequently, every time the term cell is used, it is in reference to eukaryotic cells.[4]

2.1.2 The Cell Cycle

As every living being, cells possess a life cycle. This life cycle, or the cell cycle, describes all the processes that a cell goes through to replicate its components, particularly its genome, to successfully divide into two daughter cells. [5]

Clearly, the cell cycle includes and orderly sequence of events to achieve this end and can be divided into four phases: gap 1 (G1), synthesis (S), gap 2 and mitosis (M), which occur in this order. A fifth state exists, named G0. This state corresponds to a non-dividing and resting stage and, as such, is not part of the cycle. Each of the stages possess a specific role and the entire process is controlled by a set of enzymes, the cyclin dependent kinases (CDK's). As the name sugests, these enzymes are regulated by cyclins, a protein group that appears and disappears during the cell cycle in a cyclic manner and that enables or inhibits the enzymes action.

To control the transition between phases, cells also developed a set of checkpoints along their cycle. These control the sequence and timing of the cycle, in addition to ensuring that essential events are completed successfully. As such, these checkpoints function as hold points and, in the event that a vital task is not performed correctly, they will delay the cycle's progression until the task is completed. [5]

G1 phase Along with **G0**, this stage of the cell cycle is the only one in which the cells respond to extracellular *stimuli*. As such, this phase is commonly a target for mitogenic signals (cell cycle inducing signals). At this phase, cells "decide" to enter a new round of the cellular cycle or opt to transition into a resting, quiescent state (the G0 phase). This decision is based on various intra and extra cellular signals and represents one of the previously mentioned checkpoints. If a "decision" of entering a new round is made, the cell becomes unresponsive to external signals until the end of the cellular life cycle.[6]

S phase After G1, cells enter the synthesis phase, or the **S phase**, so called due to the DNA replication which occurs during this stage, a process that starts after DNA replication proteins achieve a satisfactory level.

To ensure that DNA replication occurs in a reasonable time-frame, the process is initiated in multiple "origin points" of the chromosomes simultaneously. Nonetheless, control mechanisms have to be employed to ensure that the cells DNA content is duplicated only once and that it only restarts after the complete cell division. This is achieved by the so-called "replication licensing system".

This mechanism ensures that the thousands of "origin points" are utilized efficiently and safely, since even a small mistake during the DNA replication, be it over-replication or under-replication, could result in severe consequences for the cell. After the complete DNA replication the cell enters the Gap 2 phase. [6]

G2 Phase The G2 phase is the last step ahead of actual cellular division. As such, before progressing any further the cell must ensure that not only the genetic material is correctly duplicated, in the form of sister chromatids, but also that essential cellular structures, such as centrossomes, are too.

Incomplete DNA replication or damaged DNA will trigger checkpoint pathways that will cause cellular arrest in the G2 phase. [6]

2.1.3 M Phase

The **M phase** is the final stage of the cell cycle, after which 2 daughter cells will have been generated from 1 parent cell. This is done through two processes, mitosis and cytokinesis, that, combined, constitute the **M phase**. [6].

Mitosis is the process responsible for the separation of the cell's nucleus in two and it is comprised by 5 different stages, Prophase, Prometaphase, Metaphase, Anaphase and Telophase. Each involves characteristic steps to align and separate the cell's chromosomes.

When the last stage of mitosis has ran its course, the cell possesses two nucleus with identical genetic material and it is time to split the parent cell into two identical daughter cells, a process called **cytokinesis**.

This physical mechanism begins with the cell pinching itself at the equator forming a cleft called **cleavage furrow**. This cleft is formed by the action of a contractile ring consisting of overlapping actin and myosin filaments. As the ring tightens, it eventually reaches its smallest point. At this moment, the cell bisects itself forming two daughter cells of equal size. [7]

2.1.4 Cell Imaging

The cell cycle is a complicated process, driven and regulated by an intricate system of protein complexes that trigger specific events at specific times.[8] Being able to observe the changes in cellular structure caused by these mechanisms plays a key role in understanding how they operate. As such, the technology to observe cells needs to constantly improve producing techniques such as **fluorescent microscopy**.

Fluorescen microscopy Fluorescent microscopy is a technique that attempts to only reveal the objects of interest on an otherwise black background. To do so, it is required that the objects of interest fluoresce, which is achieved through the fluorescence phenomenon. Strictly speaking, fluorescence describes photoluminescence that occurs when materials photons at a certain wavelength and then emit photons on a different band of wavelength.

Nonetheless, the fluorescence phenomenon is only useful if the target molecules "light up". Whereas many organic substances possess natural fluorescence, the common approach from fluorescent microscopy is to use synthesized compounds that bind to a specific biological molecule and have great fluorescent properties. These compounds are known as fluorophores and provide a targeted approach since they grant the ability to only mark relevant biological molecules. [9]

DAPPI Stain One of the more relevant aforementioned fluorophores is the 4',6-diamidino-2-phenylindol or DAPPI stain. This staining compound is widely used to mark DNA with a high sensibility that allows for the observation of even small DNA quantities.

The staining process associated with this fluorophore is simple and requires no hydrolysis, with the stain being manually applied by pores made in the cellular membrane. When excited by light it emits a strong white blueish fluorescence. [10]

2.1.5 Current cell staging techniques

Cell staging techniques have changed and advanced along the years. Among the more recent ones, one is of particular relevance for this work since it provided the base biological classificaiton utilized.

Represented in [11], this method called FUCCI (fluorescent ubiquination cell cycle indicator)relies on the use of biological markers and fluorescent microscopy to identify the stage of any given cell.

Identifying antiphase oscilating proteines that mark cell-cycle transition and encoding them with fluorescent probes made visualizing a cell's current stage possible. Each color represented a different stage, cells in the G1 phase presented a red nucleus while the cells in S, G2 or M phase presented a green nucleus, as shown in figure 1.

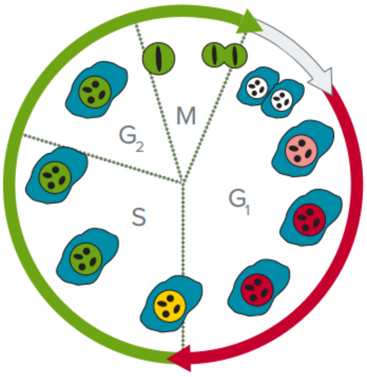


Figure 1: FUCCI staining process:Cells appear green if they are in phase G1 and red if they are in phase S/M/G2.

2.2. Machine Learning

The search for patterns in data is a fundamental problem that has a long and successful history, one that culminated in Machine Learning (ML). Machine Learning consists of classification and prediction algorithm that adjust their parameters according to the properties of a training set, giving computers the ability to learn without specific programing. [12]

ML problems can be divided into two, supervised and unsupervised problems, depending on the shape of the training set. In unsupervised learning problems, the training data consists only of input examples. It's lacking a corresponding target value that, in classification problems, corresponds to the intended class of the input. In unsupervised learning problems, a typical problem consist of finding similar examples amongst the data, a process called clustering. On the other hand, supervised learning problems possess a training set with examples of inputs and their corresponding target output. [12]

For this thesis, the training set possessed a classification for each input example, as such, supervised learning was adopted.

2.2.1 Deep Learning

Most artificial intelligence problems can be solved with the correct set of features extracted from the data and analyzed by a machine learning algorithm, the difficulty lies in choosing the relevant features and how to extract them, in other words how to represent the data in meaningful way.

Deep Learning presents a solution by introducing representations that result from automated processes. That is, deep learning provides a system built with a cascade of trainable modules. By training this system end to end, each module will adjust itself to produce the correct answer. This method allows the system to learn how to represent the raw data and how to solve the problem provided. [13]

Artificial Neural Networks Artificial Neural Networks (ANNs) are the quintessential deep learning model and receive their name from being a network of connected neurons, not unlike the brain.

Each neuron, receives inputs signals ,x_i, from various other units and computes its own output. Each input is regulated by the connection weights, w_i, which emulate biological synapses. Thanks to a transfer function, $f(z)$, the neurons possess a non-linear behavior, which is limited by a threshould β. As such, the output of a neuron can be computed as follows: [14]

$$O = tf(net) = tf(\sum_{i=1}^{n} w_i x_i + b) \tag{1}$$

Clearly, in 1 the variable net represents the scalar product of the weight vector and input vectors:

$$net = w^T z = w_1 z_1 + w_2 z_2 + ... + w_n z_n \tag{2}$$

From equation 1 it is also obvious that the transfer function will determine the neural output. The simplest case is to consider tf as a boolean step function, in which case the output can be described as:

$$O = tf(net) = \begin{cases} 1, & if w^T z > \beta \\ 0, & else \end{cases} \tag{3}$$

In addition to the transfer function, the connecting weights also determine the output value. It is by adjusting the weight vector that each neuron is capable of learning. Since the neurons are connected amongst themselves in a network, this learning ability is inherited by the network. Artificial neural networks are organized in distinct layers. The input layer receives the network's input vector while the output layer produces the intended output. Between these two layers, a model can have multiple hidden layers, depending on its depth. Each neuron connects only to the neurons in the previous and next layer, each connection having a respective connecting weight. To adjust these weights, effectively training the network, a common method is to use back propagation [15]

The weight update rule can be defined as:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \sum \alpha \Delta w_{ij}(m-1) \tag{4}$$

Where η corresponds to the learning rate which determines the rate of change in the networks weight and α corresponds to the momentum which determines the effect the past $m-1$ weight changes on the current direction of movement in the weight space. E corresponds to the error function, here the mean squared error function was used:

$$E = \frac{1}{2} \sum_{i=1}^{n} (O_{ij} - T_{ij})^2 \tag{5}$$

Where O is the network's output vector and T is the target output vector for a specific input.

As expression 4 clearly demonstrates, the appropriate selection of the learning rate and the momentum plays a crucial part in both the speed and success of the training.

2.2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) present themselves as an evolution of "traditional" neural networks. Essentially, CNNs are neural networks that use possess at least one convolutional layer. That is, at least one layer that uses the mathematical convolution operation instead of a general matrix multiplication. [16] Mathematically, the convolution expresses the amount of overlap a function g produces as it is shifted over another function f, the product function being a "blend" of the two.

Commonly, convolution is noted using an asterisk operator. As such it can be expressed as follows:

$$[f * g](\tau) = \int f(\tau) g(t-\tau) d\tau \tag{6}$$

When applying convolution to machine learning, the first argument (f in equation 6) is called input, the second (g in equation 6) is called kernel and the output is commonly called feature space. Also, in machine learning the discreet definition of convolution is used:

$$[f * g](\tau) = \sum_{\tau=-\inf}^{\inf} f(\tau)g(t-\tau) \tag{7}$$

In most CNN models, the input is multi-dimensional and the kernel is a multi-dimensional array of parameters tuned by the learning method. Figure 2, gives a visual representation of how the convolution operations works in 2-D arrays, as is the case of images.

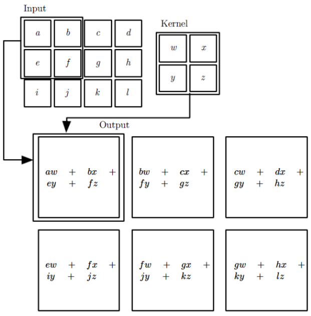


Figure 2: An example of a 2-D convolution without kernel flipping. Source:[14]

In addition to the convolutional operation, the convolutional layer of a network has two more stages. A detector stage, where each linear activation is transformed by a non-linear function and the pooling stage, where a pooling function further alters the output. Pooling functions replace the output with the a summary statistic of neighboring outputs, merging semantically similar features into one.

Convolutional neural networks have enjoyed relative success in detecting, segmenting and recognizing objects since the early 2000's. Nonetheless, they were largely forgotten by the computer vision community until the ImageNet competition. In this competition (which consists in using machine learning algorithms to classify a large image dataset in one thousand classes) CNNs performed remarkably, achieving extraordinary results and thus becoming the standard approach for computer vision problems. [17]

3. Implementation

In this section a brief explanation on the processed biological data that was the basis of the work performed in this thesis will be given. Most of the work was performed prior to this master thesis and provided the initial data utilized during this project.

Additionally, an explanation of the deep learning methods utilized during this project will be provided. All of the artificial intelligence methods were implemented in Python utilizing the Keras and Tensorflow libraries.

3.1. Biological Material and Data

3.1.1 Cell culture and imaging

Even tough only the DAPI stain information was utilized to identify the features that reflect the cells progress along the cell cycle, fluorescent imaging provided the basis for this work. In total, 836 images were obtained from 12 samples comprising 5873 cells from NMuMG-Fucci2 *in-vitro* cultures were obtained from Riken institute Japan.

To extract all the information contained of the DAPI stained nuclei, multiple images were taken along the z axis and merged together by projecting into a single image.

3.1.2 Image Pipeline

The images obtained from the fluorescent imaging were then processed to produce the data utilized as starting point for this thesis. The preprocessing work was performed before this project and a brief explanation will be provided based on [18].

The image preprocessing pipeline consists of two steps, an initial application of a denoising algorithm and contrast/ intensity adjustments followed by the segmentation of each nucleous.

Nuclei plane image denoising Denoising is utilized in image processing to reduce the effect of noise generated by the instrumentation systems into the samples and to emphasize the underlying relevant data.

The noise introduced in fluorescent microscopy follows the poison distribution, with a probability function that can be defined as:

$$Pr(X = k) = \frac{\lambda^k exp^{-k}}{k!} \tag{8}$$

where λ is the distribution parameter.

A Bayesian algorithm was employed to remove the noise with maximum-a-posterior optimization criterion:

$$\hat{Z} = argmin_z E(Z,Y) = argmin_z(E_y(Z,Y) + E_z(Z)) \tag{9}$$

where $E_y(Z,Y)$ is the data fidelity term and $E_z(Z)$ the prior term regularizing the solutionm required to introduce some *apriori* information about the solution. Assuming observations are independent and the noise compliant with a Poisson distribution, the data can be described as:[19]

$$E_y(Z,Y) = -\log\left[\Pi_{i,j=0}^{N-1,M-1} p(Y_{i,j}|z_{i,j})\right] = \sum_{i,j=0}^{N-1,M-1} |z_{i,j}y_{i,j}\log(z_{i,j} \tag{10}$$

where C is constant.

The distribution in 10 is denominated by a log total variation potential function:

$$TV \log = \sqrt{\log^2 \frac{z}{\varsigma}} \tag{11}$$

where z and ς are neighboring pixels.

The function described in 11 has efficient high frequency noise removal in homogeneous regions and a smaller penalization in sharp transitions, useful for biological images who possess abrupt transitions that are desired.

Segmentation With the segmentation process, the cell culture images were divided into multiple images, each containing a nucleous of a cell. This process has a typically low accuracy and inconsistent output when applied to most images, and is crucial to ensure the success of the final analysis.

For this data, the segmentation strategy was utilized in the denoised and contrast/ intensity DAPI plane (blue channel) of the FM images, and consisted in the application of Otsu thresholding and morphological operators to each image [18].

3.2. Deep learning

To determine the cellular stage of each cell and have the ability to classify each nucleous independently (i.e. not rely in clustering methods), machine learning algorithms were employed. Namely artificial neural networks (ANN) and convolutional neural networks (CNN).

3.2.1 Neural Networks

Feature extraction The first step in using an ANN to learn the cellular stage, was to select and extract relevant features from the dataset. The features identified were the nucleous area and total intensity. which reflect the changes that happen to a cell nucleous during the cellular cycle.

Each feature was expressed mathematically. The area was defined as the total number of pixels contained in a nucleous (NP) such as:

$$Area = \sum n_{pn} \tag{12}$$

The total intensity was defined as the sum of the intensity of each pixel in the nucleous (as shown below) and, theoretically, reflects the amount of DNA contained in each cell.

$$Area = \int_A intensity dA = \sum_{i=1}^{N} Intensity \tag{13}$$

Data normalization After extracting the features, data normalization was applied to the data set, namely, z-score normalization.

Data normalization is a common pre-processing technique to minimize the impact features with different ranges have in the final outcome. Potentially, features with larger ranges would have a higher contribution for the final outcome than the features expressed in a smaller range. Furthermore attributes should be dimensionless so that the unit of measure does not impact the final output. As previously stated the method utilized was the z-score normalization, which can be defined as [20]:

$$x^*_{ij} = \frac{x_{ij} - \tau_{ij}}{\sigma_j} \tag{14}$$

where x^*_{ij} is the normalized attribute value, x_{ij} represents the raw data and τ_j and σ represent the mean and standard deviation (STD) for the values of the j^th attribute. Z-score normalization return with 0 average and standard deviation of 1 and is the most commonly used employed standardization technique. [21]

Designing and training the ANN With the features extracted and the visual classification from the FUCCI method, a supervised learning method was employed to train an ANN.

To size the required network it was first created a network that could "memorize" the problem, i.e. overfit the dataset, and then over-fitting prevention methods were applied. For this, the drop-out technique was employed.

In this method the term dropout refers to "dropping" units in a neural network along with incoming and outgoind connections. The choice of which figures to drop is random, with a fixed independent probability of p for each unit to drop. p can be selected form a test dataset, however a probability of 50%, i.e. p=0.5 seems to be optimal for most neural network problems [22].

By dropping units, essentially a thinned network is trained each time and is formed by all the remaining units. As such, training a network with elements can be seen as training a set of 2^n thinned networks with weight sharing.

A good approximation for the resulting thinned network is to use a neural network without the dropout where the weights of each hidden unit is multiplied by p at validation time [22].

To adjust the training of the designed network 3 parameters were adjusted, nuber of epochs, batch size and learning rate. These parameters are correlated among them, for example, smaller learning rates require more training epochs since each update has a smaller effect on the weights. On the other hand, batch size determines the number of times the error function is determined per epoch and subsequently the number of times the model weights are updated in each epoch.

3.2.2 Convolutional Neural Networks

Image generation Contrary to ANN's, CNNs are capable of feature detection and selection. As such, instead of extracting the features, as done for the ANN method, 150×150 pixel images were produced from the information generated by the preprocessing pipeline previously explained in this section.

Data augmentation The use of CNNs for image classification requires a large amount of data to prevent over-fitting. Unfortunately, the original data set only presented 5873 nucleous and, as such, data augmentation, a data-space solution to the problem of limited information for deep learning, was employed during this work.[23].

While not the only solution to overfitting in deep neural networks, data augmentation addresses the problem at its root, the training data set by assuming more information can be extracted from the original data through augmentation. To extract this additional information, data augmentation methods typically inflate the training dataset size by either data warping or oversampling, with techniques from geometric and color transformations to random erasing, adversarial training or neural style transfer.

During this project, only data warping was employed. However, the main goal of data augmentation is to inflate the available data while still keeping the label of each data point valid. Since the two most relevant features for describing the progression in the cell cycle are area and intensity a careful selection of the techniques employed is required to ensure these features are not perturbed and negatively affect the outcome. After some initial testing, the best results were achieved by utilizing only rotation, flipping and translations in the training and in the test set.

Designing and training the CNN While data augmentation prevents overfitting and helps the network ”learn” more robust features, it is not the only overfitting prevention method available. In addition to the dropout method previously described, early stopping was utilized to ensure the designed network would be perform optimally. This method consists of monitoring the validation accuracy and loss to detect over fitting and training is then stopped before convergence.

3.3. Validation

To validate the training and designing methods explained previously, both the ANN and CNN network were tested against the modified National Institute of Standards and Technology (MNIST) data set [24]. This dataset contains images from handwritten digits and is one of the most common starting points for neural networks with the performance of various types of networks well documented.Testing both the ANN and CNN networks designed for this project with this data set, a validation error of ¡ 8% was achieved for both of them.

In addition to using the MNIST data set for validation, a pre-trained VGG [25] network was as a benchmark for the convolutional neural network.

4. Results

In this section, the results obtained during this work are presented. First, the results for the aritificial neural network are shown with a demonstration of the over fitting prevention methods followed by the demonstration of the results obtained utilizing the convolutional neural network.

4.1. Artificial Neural Network

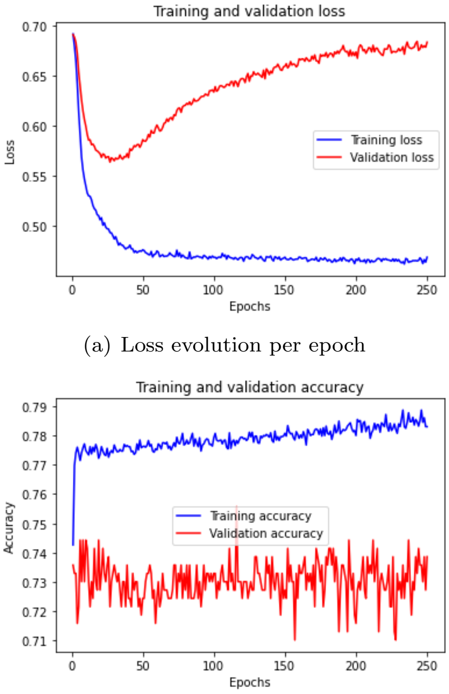
After extracting the selected features from the biological dataset, area and intensity, these were subjected to classification using an artificial network. The designed network was capable of over fitting the problem, evidenced by the separation between the validation loss and the training loss represented in figure 3(a). Figure 3(b) represents the accuracy evolution of the same training, and while the initial weights loaded into the network provided a good starting point, without over fitting prevention methods, the validation accuracy was unstable and it did not increase, showing the models inability to generalize the information acquired during training

Figure 4 shows the training evolution of the same network with over-fitting prevention methods utilized, namely the use of dropout. As is evidenced by 4(a), in this case, while not exactly converging , the training loss and the validation loss do not diverge as the network is trained with stabilization of the accuracy validation around 70%

To simulate early stopping, the best performing network was saved and then tested against a validation set that was not used during training and evaluated across 3 metrics: sensitivity, specificity and accuracy.

$Senstivity$	$Specificity$	$Accuracy$
73,13%	78,81%	76,21%

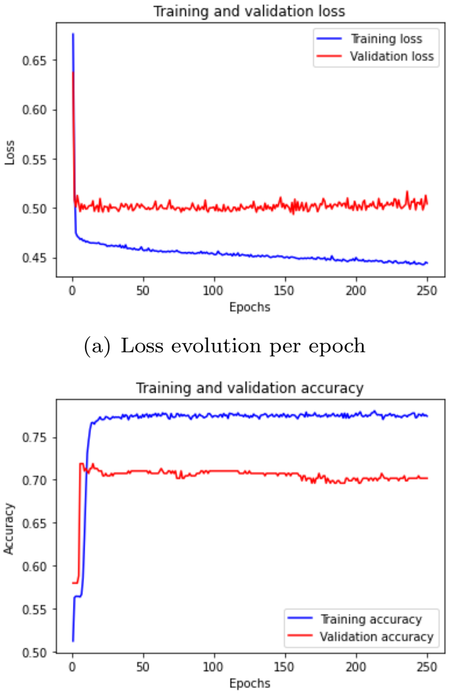
Table 1: ANN validation results.



(a) Loss evolution per epoch

(b) Accuracy evolution per epoch

Figure 3: Training and validation metrics for ANN without dropout method.



(a) Loss evolution per epoch

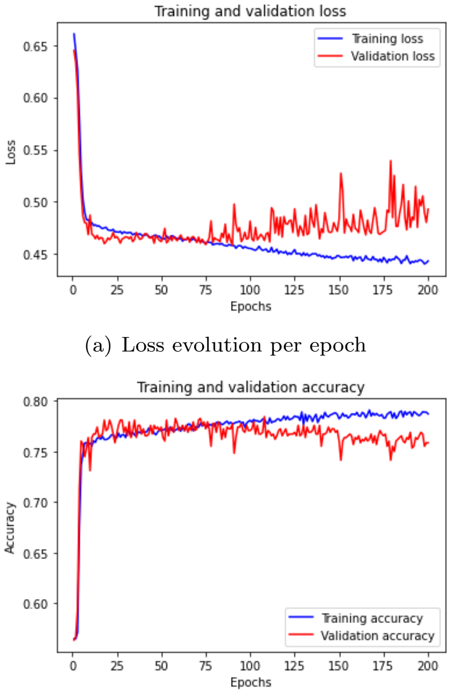
(b) Accuracy evolution per epoch

Figure 4: Training and validation metrics for ANN with dropout method.

4.2. Convolutional Neural Network

On this section, the results obtained from the use of convolutional neural networks will be demonstrated.

Figure 5 represents the evolution of the CNN metrics during training. As shown in figure 5(a), the loss started reducing, however from approximately epoch 75 onwards it is clear the training is resulting in some over fitting. This is also evidenced in figure 5(b) where the validation accuracy starts decreasing while training accuracy evolves. As such, training was interrupted and the parameters were fine tuned to improve the performance of the network. Figure 6, shows the evolution of this training.



(a) Loss evolution per epoch

(b) Accuracy evolution per epoch

Figure 5: Training and validation metrics for CNN trained with synthetic data.

Once again, the weights of the best performing CNN were utilized to assess the network performance across the same metrics as the ANN. Below are the results:

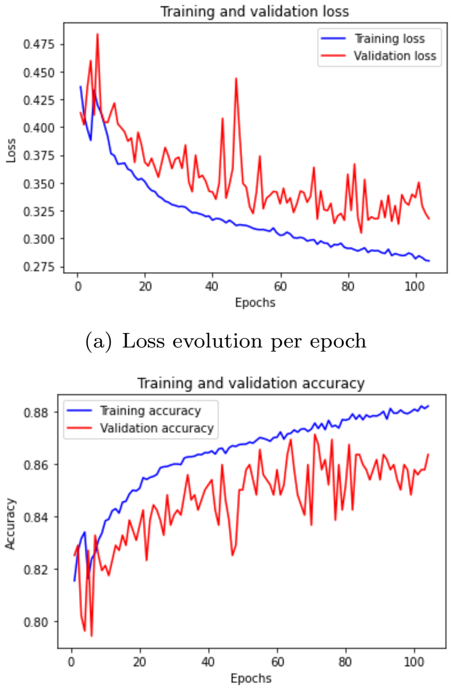
Senstivity	*Specificity*	*Accuracy*
78,31%	90,88%	86,93%

Table 2: CNN validation results.

As stated previously, the VGG network was utilized as a benchmark for the CNN performance and was trained on the same data set with the same data augmentation conditions of the designed CNN. Below are the results:

$Senstivity$	$Specificity$	$Accuracy$
98,23%	73,17%	82,95 %

Table 3: VGG validation results.



(a) Loss evolution per epoch

(b) Accuracy evolution per epoch

Figure 6: Training and validation metrics for CNN fin tuned with synthetic data.

5. Conclusions

The following section will present the achievements of this thesis and delineate a path for future work.

Based on FM images obtained of *in vivo* cell cultures, the primary goal of this thesis was to develop a simple way of correctly identifying the cell phase of a particular cell.

Currently, FM based methods for accessing the cell status of individual cells are only capable of probing specific parts of the cell cycle [26], or are very laborious since they evolve the growth of cultures with phase specific identifiers, such as the FUCCI method utilized as validation for this work. While these second methods are capable of monitoring all phases in the cell cycle, this approach results in a very complex process which requires the use of multiple imaging channels and inhibit the capability to visualize other cell features in the same culture.

By contrast, the method suggested in this thesis relies on the use of a inexpensive and commonly used compound, the DNA dying die DAPI. This compound, allows the extraction of information representing both the nucleus area and the amount of DNA, two intrinsic features who provide a good description of the cell status along its cellular cycle.

While some other machine learning methods that really on the information gathered from the DAPI stain have already been proposed [18], they relied on the use of clustering algorithms which not capable of classifying a single data point. The method proposed for this work implemented the use of deep learning techniques, were capable of identifying the cell phase from a single data point.

The ANN method relied on the manual extraction of the intrinsic features of the nucleus, area and intensity, that should translate adequately the cell phase of a single nucleus. By utilizing this method an accuracy of $\approx$ 76% was obtained when classifying cells in either the G1 or the S/G2/M phase. A result which, while lower than the accuracy obtained in clustering methods, still demonstrates a clear correlation between the selected features and the cell phase.

In addition to the ANN method, two CNNs were also utilized in this thesis in conjunction with data augmentation techniques. The first network was a self designed simple CNN with two convulational layers which produced an accuracy of $\approx$ 87% while the second network utilized resorted to the architecture of a VGG netowrk which produced an accuracy of $\approx$ 83%. Both of these networks relied on the automatic identification of features that translated the cell phase. While the results achieved are still below the ones obtained with clustering algorithms, the increase in accuracy compared with the ANN still demonstrates that the use of a CNN to adress image based cell classification is a promising solution.

While below the results achieved in previous work, the results achieved with the methods proposed were very satisfactory and represent a promising solution to the problem of image based cell classification for a single nucleus. However, a more complex CNN should be utilized and fine tuned with a larger dataset than the one currently available to try and improve the results obtained during this work.

References

[1] Dean Jackson and Peter R. Cook. Analyzing dna replication i: Labeling animals, tissues, and cells with bromodeoxyuridine (brdu). *Cold Spring Harbor Protocols*, 2008(8):pdb.prot5031, 2008.

[2] Michael Hesse, Alexandra Raulf, Gregor-Alexander Pilz, Christian Haberlandt, Alexandra M. Klein, Wilhelm Röll, Michael I. Kotlikoff, Christian Steinhäuser, Magdalena Götz, Hans R. Schöler, and Bernd K. Fleischmann. Direct visualization of cell division using high-resolution imaging of m-phase of the cell cycle. *Nature Communications*, 3:1076, 2012.

[3] Gianlucaand Voss Ty C.and Misteli Tom Roukos, Vassilisand Pegoraro. Cell cycle staging of individual cells by fluorescence microscopy. 10:334–348, 2015.

[4] C. M. O'Connor and Jill U. Adams. *Essentials of Cell Biology.* NPG Education, 2010.

[5] C. M. O'Connor and Jill U. Adams. *Essentials of Cell Biology.* NPG Education, 2010.

[6] Mathewos Tessema, Ulrich Lehmann, and Hans Kreipe. Cell cycle and no end. *Virchows Archive*, 444(4):313–323, 2004.

[7] C. M. O'Connor and Jill U. Adams. *Essentials of Cell Biology.* NPG Education, 2010.

[8] Xavier Graña and E. Premkumar Reddy. Cell cycle control in mammalian cells: role of cyclins, cyclin dependent kinases (cdks) growth supressor genes and cyclin-dependent kinase inhibitors (ckis). *Oncogene*, 11:211–219, 1995.

[9] Jeff W Lichtman and José-Angel Conchello. Fluorescence microscopy. *Nature Methods*, 2(12):910–919, 2005.

[10] Betty I. Tarnowski, Francis G. Spinale, and James H. Nicholson. Dapi as a useful stain for nuclear quantitation. *Biotechnic & Histochemistry*, 66(6):296–302, 1991.

[11] Asako Sakaue-Sawano, Hiroshi Kurokawa, Toshifumi Morimura, Aki Hanyu, Hiroshi Hama, Hatsuki Osawa, Saori Kashiwagi, Kiyoko Fukami, Takaki Miyata, Hiroyuki Miyoshi, Takeshi Imamura, Masaharu Ogawa, Hisao Masai, and Atsushi Miyawaki. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, 132(3):487 – 498, 2008.

[12] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, chapter 1, pages 1–12. Springer-Verlag, 2006.

[13] Yann LeCun. The power and limits of deep learning: In his iri medal address, yann lecun maps the development of machine learning techniques and suggests what the future may hold. *Research-Technology Management*, 61(6):22–27, 2018.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, chapter 6, pages 168–227. The MIT Press, 2016.

[15] Ajith Abraham. *Handbook of Measuring System Design*, chapter Artificial Neural Networks. John Wiley & Sons, Ltd., 2005.

[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*, chapter 9, pages 330–372. The MIT Press, 2016.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[18] Ivan Sahumbaiev. Cell cycle staging from dapi and fluorescence microscopy. Master's thesis, Instituto Superior Técnico, 2015.

[19] I. C. Rodrigues and J. M. R. Sanches. Convex total variation denoising of poisson fluorescence confocal images with anisotropic filtering. *IEEE Transactions on Image Processing*, 20(1):146–160, 2011.

[20] Mikhail Y. Prostov Maria M. Suarez-Alvarez, Duc-Truong Pham and Yuriy I. Prostov. Statistical approach to normalization of feature vectors and clustering of mixed datasets. *The Royal Society*, 2012.

[21] Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data clustering: theory, algorithms, and applications.* SIAM, 2007.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[23] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, Jul 2019.

[24] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[26] Dean Jackson and Peter R Cook. Analyzing dna replication i: labeling animals, tissues, and cells with bromodeoxyuridine (brdu). *Cold Spring Harbor Protocols*, 2008(8):pdb–prot5031, 2008.