scientific reports

OPEN

A machine learning approach for single cell interphase cell cycle staging

Hemaxi Narotamo[1,6], Maria Sofia Fernandes[2,3,6], Ana Margarida Moreira[2,3,4], Soraia Melo[2,3], Raquel Seruca[2,3,5]✉, Margarida Silveira[1] & João Miguel Sanches[1]

The cell nucleus is a tightly regulated organelle and its architectural structure is dynamically orchestrated to maintain normal cell function. Indeed, fluctuations in nuclear size and shape are known to occur during the cell cycle and alterations in nuclear morphology are also hallmarks of many diseases including cancer. Regrettably, automated reliable tools for cell cycle staging at single cell level using in situ images are still limited. It is therefore urgent to establish accurate strategies combining bioimaging with high-content image analysis for a bona fide classification. In this study we developed a supervised machine learning method for interphase cell cycle staging of individual adherent cells using in situ fluorescence images of nuclei stained with DAPI. A Support Vector Machine (SVM) classifier operated over normalized nuclear features using more than 3500 DAPI stained nuclei. Molecular ground truth labels were obtained by automatic image processing using fluorescent ubiquitination-based cell cycle indicator (Fucci) technology. An average F1-Score of 87.7% was achieved with this framework. Furthermore, the method was validated on distinct cell types reaching recall values higher than 89%. Our method is a robust approach to identify cells in G_1 or S/G_2 at the individual level, with implications in research and clinical applications.

The cell cycle is a highly organized and coordinated process that ensures the correct duplication of genetic material and cell division[1,2]. Importantly, the progression of cells through the cell cycle occurs in an orderly sequence of events and encompasses four distinct cell cycle phases termed G_1 (Gap1), S (synthesis), G_2 (Gap2) and M phase or mitosis[2,3]. Briefly, in each cell cycle, cells in G_1 prepare for DNA synthesis, which occurs in the S phase, and subsequently progress to G_2 to prepare for mitosis[2,3]. All phases are tightly regulated by cell cycle checkpoints, including cyclins and cyclin dependent kinases (CDKs), that control and ensure cells are able to proceed along the cell cycle[1,4].

Remarkably, dysregulation of the cell cycle occurs in many diseases, such as cancer[1,5]. Indeed, it has been shown that tumor cell cycle dynamics has prognostic value in many cancer types including breast, gastric and prostate cancer[6–8]. Furthermore, cell cycle evaluation has been reported to predict sensitivity or resistance to chemotherapeutic regimens and specific therapeutic strategies for various cancers[9,10]. In addition, cell cycle proteins and regulators have become attractive targets in cancer therapy and novel cell cycle inhibitors have emerged to provide new treatment options for cancer patients[1,11]. Thus, determining cell cycle phases is of critical importance for tumor characterization and monitoring, ultimately impacting cancer care.

To date, most studies on cell cycle staging involve flow cytometry or other methods that often use cells in suspension, analyze cell populations or require specific cell cycle markers and extensive cell manipulation, thus presenting many drawbacks[12]. One example of cell cycle assessment requiring specific cell cycle markers is the recent fluorescent ubiquitination-based cell cycle indicator (Fucci) technology. Fucci is a genetically encoded indicator system of cell cycle progression, in which cells are modified to express cell cycle markers[13–15]. More specifically, the Fucci technology is a fluorescent protein based sensor system that takes advantage of Cdt1 and Geminin, two proteins that oscillate inversely and are involved in the DNA replication control system. These are fused to red and green fluorescent proteins allowing the identification of cells at G_1 and $S/G_2/M$, with nuclei in G_1 phase red and nuclei in $S/G_2/M$ phases green[13–15].

[1]Institute for Systems and Robotics (ISR), Instituto Superior Técnico (IST), University of Lisbon, Lisbon, Portugal. [2]Epithelial Interactions in Cancer (EPIC) Group, Instituto de Investigação e Inovação em Saúde (i3S), University of Porto, Porto, Portugal. [3]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), University of Porto, Porto, Portugal. [4]Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Porto, Portugal. [5]Faculty of Medicine, University of Porto, Porto, Portugal. [6]These authors contributed equally: Hemaxi Narotamo and Maria Sofia Fernandes. ✉email: rseruca@ipatimup.pt

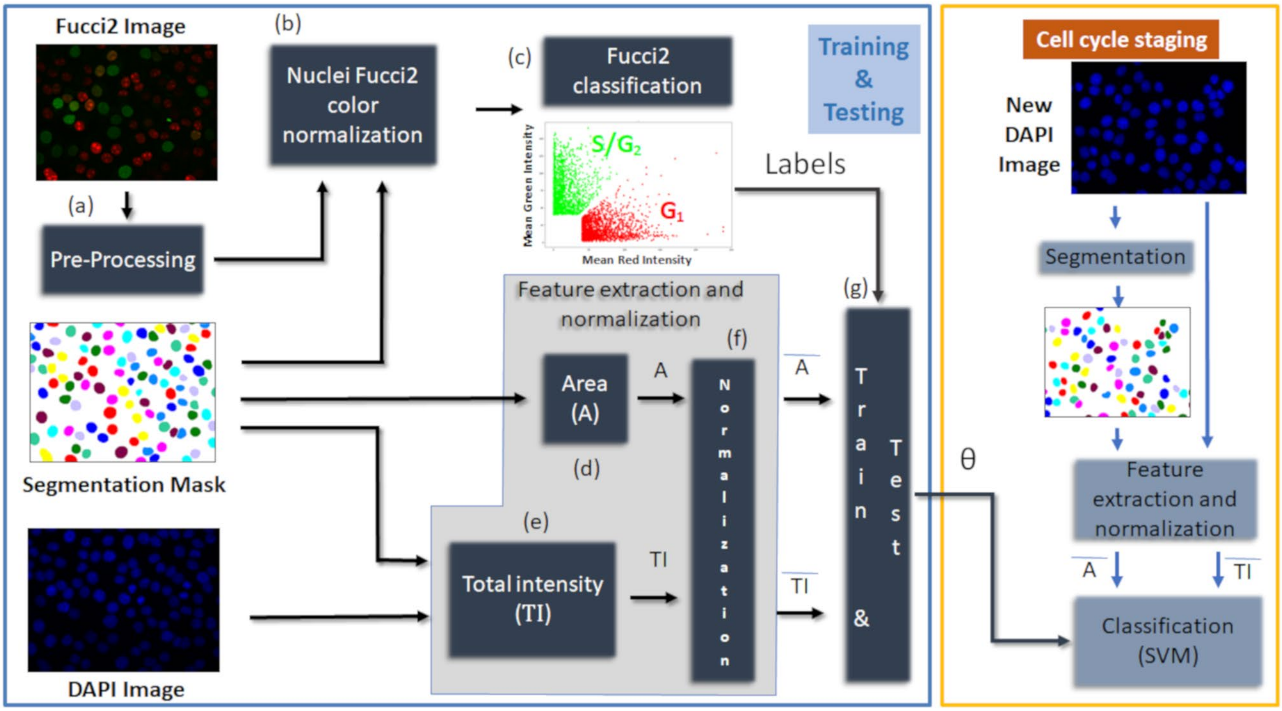


Figure 1. Overall pipeline to train and test the SVM for interphase cell cycle staging. The processing pipeline involves the analysis of DAPI and Fucci2 images, which are required to train and test the classifier (blue box, **a–g**). The final single cell cycle staging procedure uses the classifier parameters, obtained in the training phase, to identify the cell cycle phase of new DAPI images (yellow box); *SVM* support vector machine, *Fucci* fluorescent ubiquitination-based cell cycle indicator.

In recent years, novel approaches are emerging taking advantage of nuclear morphological alterations occurring along the cell cycle[16], thus opening new opportunities for the development of automated bioimaging methods for cell cycle classification. Several strategies based on image processing have been proposed to infer the cell cycle phase of cells based on their DNA content, shape and texture[17–20]. Regrettably, most of the methodologies do not focus on the interphase phases G_1, S and G_2, since these are more difficult to classify at cell level or present limitations mainly related to manual parameter tuning[17–22].

Thus, there is an urgent need to develop automated tools for in situ interphase cell cycle staging of single cells within heterogeneous cell populations as cancer cells, using microscopy images and taking advantage of automated image analysis. In the present work, we propose a combined strategy using deep learning followed by a supervised classifier for interphase cell cycle staging of individual cells using in situ fluorescence images of nuclei stained with the DNA dye 4′,6-diamidino-2-phenylindole (DAPI). Specifically, we present a new pipeline using a deep learning cell nuclei segmentation method[23] and involving training and testing a machine learning classifier for categorization of nuclei patches from DAPI images in G_1 or in S/G_2 classes. Noteworthy, for training, nuclei labels were obtained with molecular biomarkers according to the Fucci technology using Fucci2, a Fucci derivative with different fluorescent properties[13,15]. Subsequently, to further validate our strategy, cell cycle staging was performed on a panel of human gastric, breast and colorectal cancer cell lines. Overall, our data strongly indicates that this method is a robust strategy to successfully identify cells in situ in G_1 or S/G_2 at the individual level from distinct cell types, in a rapid and effective manner, which can be applied to heterogeneous cell populations such as cancer cells.

Results

In this work, we developed a new analytical bioimaging pipeline for interphase cell cycle staging of individual cells using in situ fluorescence images. The strategy involved segmentation, training, cross-validation and testing of a classifier to ensure proper categorization of new data. The overall pipeline is schematically represented in Fig. 1. In particular, the staging is performed by a *Support Vector Machine* (SVM) classifier based on features extracted from nuclei stained with DAPI. The labels, required to train it, are automatically computed from the corresponding image planes of cells expressing Fucci2 fluorescent cell cycle probes using the Fucci technology, available in this dataset for training and performance evaluation. Ultimately, the result of our training procedure is a set of parameters θ_{SVM}, which can be used to classify new DAPI nuclei patches regarding their cell cycle phase based solely on DAPI staining.

As depicted in Fig. 1, training and testing of the SVM classifier included:

Feature	Equation	
Area: number of pixels in the nucleus	$A_k = \sum_{i,j} b_k(i,j)$	(1)
Total DAPI intensity: sum of the intensities of the pixels in the nucleus extracted from the DAPI image	$TI_k^b = \sum_{i,j} b_k\left(i,j\right)D(i,j)$	(2)
Total red intensity: sum of the red channel's intensities of the pixels in the nucleus extracted from the Fucci2 image	$TI_k^r = \sum_{i,j} b_k\left(i,j\right)F^r(i,j)$	(3)
Total green intensity: sum of the green channel's intensities of the pixels in the nucleus extracted from the Fucci2 image	$TI_k^g = \sum_{i,j} b_k\left(i,j\right)F^g(i,j)$	(4)
Mean red intensity	$\mu_k^r = \frac{TI_k^r}{A_k}$	(5)
Mean green intensity	$\mu_k^g = \frac{TI_k^g}{A_k}$	(6)
Normalized red intensity: to capture the red color independently of the intensity	$\overline{\mu}_k^r = \frac{\mu_k^r}{\sqrt{(\mu_k^r)^2+\left(\mu_k^g\right)^2}}$	(7)
Normalized green intensity: to capture the green color independently of the intensity	$\overline{\mu}_k^g = \frac{\mu_k^g}{\sqrt{(\mu_k^r)^2+\left(\mu_k^g\right)^2}}$	(8)

Table 1. Intensity and morphological features equations.

1. Fucci2 data processing: (a) image pre-processing; (b) color normalization based on the nuclei segmentation masks and (c) binary automatic classification assuming linearly separable classes in the red-green color space.
2. Features computation: (d) nuclei areas (A) computation from the nuclei segmentation masks, and (e) computation of the total intensity (TI) of nuclei stained with DAPI using the nuclei segmentation masks.
3. Features normalization, training and testing: (f) Features (A, TI) normalization, and (g) SVM classifier training and testing using normalized features and labels obtained in (c) from Fucci2 data. The resulting SVM parameters θ are used to classify new DAPI nuclei regardless Fucci2 information.

For classification of new DAPI images, nuclei are segmented using the algorithm developed by Narotamo et al.[23] and features (A, TI) are extracted and normalized as in (d), (e) and (f). The classifier parameters obtained in (g), θ_{SVM}, are then used to classify new images.

Automatic classification of Fucci2 cell cycle labels required for machine learning training.

In the present study, a supervised approach for interphase cell cycle staging was developed. For this purpose, supervision and training of the machine learning classifier was performed using cell cycle molecular labels generated based on Fucci technology and automatic image processing (Fig. S1).

Fucci2 image processing. In order to train the classifier, processing of the Fucci2 images was necessary. The image processing included an initial pre-processing for background removal, nuclei segmentation, feature extraction and color normalization.

More specifically, a pre-processing framework was applied to Fucci2 images so the intensities of all images were comparable, thus accounting for the variability inherent to immunofluorescence and image acquisition (Fig. S1a). In this step, for each color channel of each image, the mean background intensity was subtracted as described in Materials and Methods. Next, in the segmentation step, we aimed to clearly and efficiently identify and isolate nuclei from the images. A deep learning based segmentation method proposed by Narotamo et al.[23] was used, in which nuclei were first detected through the Fast YOLO architecture, as detailed in Materials and Methods. The resulting masks, each one containing a single nucleus, were used to compute the labels from the Fucci2 images required for classification (Fig. S1b). Finally, for feature extraction and color normalization, the features from the kth nucleus were computed from the original images by multiplying DAPI, $D(i, j)$ or Fucci2, $[F^r(i, j), F^g(i, j)]$ by the binary mask associated with that nucleus, $b_k(i, j)$, obtained in the segmentation step. The resulting nucleus specific intensity images are $d_k(i, j) = b_k(i, j)D(i, j)$ and $[f_k^r(i, j), f_k^g(i, j)] = b_k(i, j)[F^r(i, j), F^g(i, j)]$. Finally, for all the other nuclei, the extracted features were computed according to the Eqs. (1)–(4) (Table 1). Based on these features, additional parameters were computed according to the Eqs. (5)–(8) (Table 1).

Establishment of Labels from Fucci2 Data. Fucci2 data was used to automatically compute cell cycle labels of each nucleus required to train the classifier. Each kth nucleus is represented by a single colored dot in the two-dimensional (2D) space of red–green (RG) colors by its mean intensities $[\mu_k^r, \mu_k^g]$, according to Eqs. (5) and (6) as shown in Fig. 2a. The RGB color of each dot is $[\mu_k^r, \mu_k^g, 0]$. Normalized values are represented in Fig. 2b, in which the color of each dot is computed using the Eqs. (7) and (8) to improve the ability to discriminate the three classes G_1, S and G_2. The goal of this normalization is to obtain the true color of each dot independently of its intensity. It is the main added value of the proposed automatic approach for label classification when compared with the manual, in which nuclei true color identification is difficult when intensities are low.

In Fig. 2b, we can observe each nucleus with the normalized intensity and the unit slope straight line (represented in blue) that separates G_1 and G_2 according to Fucci2 technology. At transition, the technique is not conclusive because cells are undergoing from G_1 to S. The normalization procedure is able to enhance color differences between both classes by normalizing to 1 the norm of the color vector of each nucleus

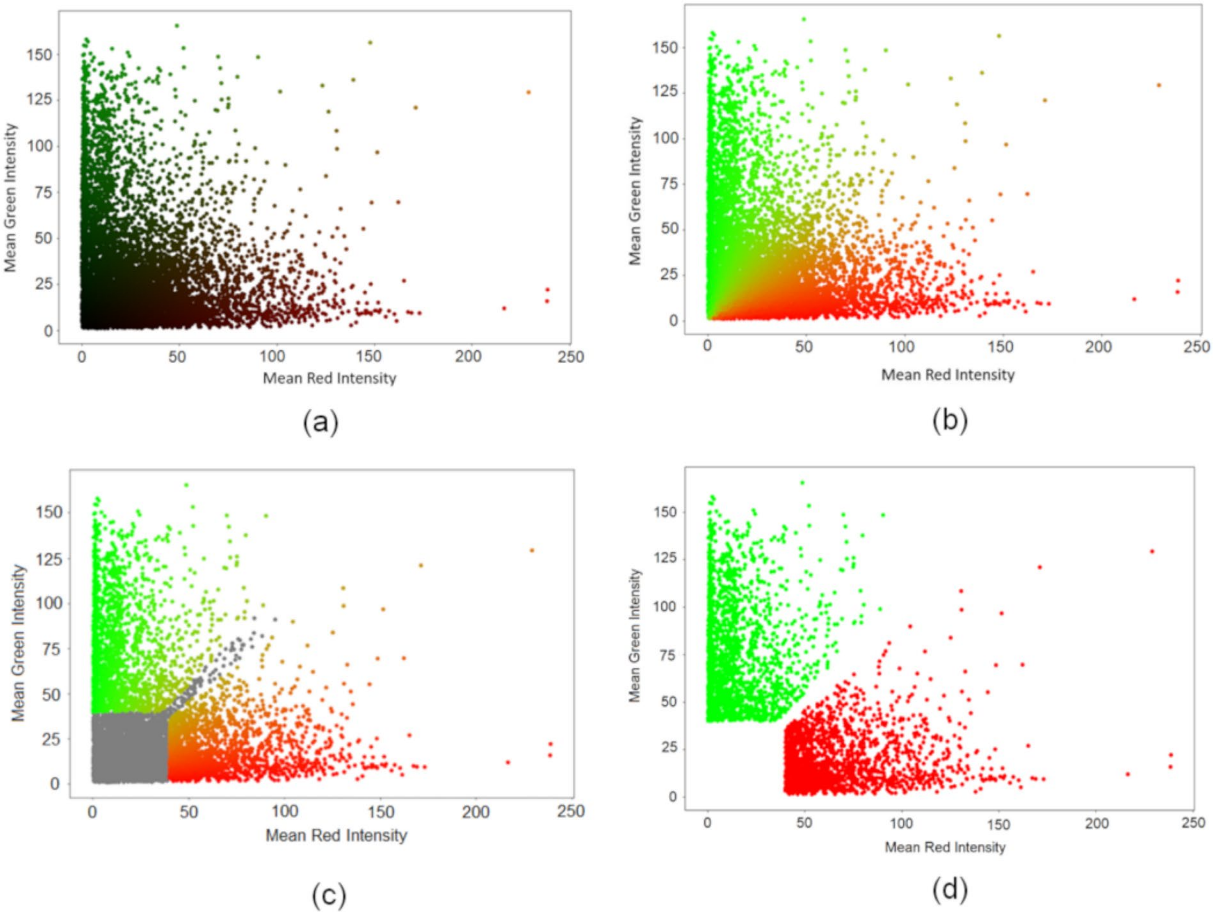


Figure 2. Nuclei distribution in the 2D space $[\mu_k^r,\mu_k^g]$ and strategy to automatically estimate the labels from the Fucci2 data to train the supervised classifier. (**a**) Each point is colored with the intensities (R,G,B) = $[\mu_k^r, \mu_k^g, 0]$. (**b**) Each point is colored with the intensities (R, G, B) $=\left[\overline{\mu}_k^r, \overline{\mu}_k^g, 0\right]_k$. (**c**) The nuclei that were excluded are colored in grey. All the other nuclei are colored with the following intensities (R, G, B) = $\left[\overline{\mu}_k^r, \overline{\mu}_k^g, 0\right]$. (**d**) Representation of each nucleus in the 2D space $[\mu_k^r,\mu_k^g]$ after removing the colorless nuclei, outliers, and some nuclei in the transition between red and green. Each nucleus is represented by a single colored dot in the 2D space. Each point is colored with its label, i.e., nuclei labeled as S/G$_2$ are green, whereas nuclei labeled as G$_1$ are red.

$\left[\overline{\mu}_k^r, \overline{\mu}_k^g\right] = \frac{\left[\mu_k^r,\mu_k^g\right]}{\sqrt{(\mu_k^r)^2+\left(\mu_k^g\right)^2}}$ and low intensity nuclei become clearly visible. However, since no additional information was added, the confidence on these nuclei is still reduced.

Thus, low intensity nuclei and nuclei at the transitions or aberrant areas were excluded from the analysis as shown in Fig. 2c (in grey) according to the following criteria, in which the parameters were chosen in a trial and error basis:

1. $\mu_k^r,\mu_k^g < \varepsilon$, too dark ($\varepsilon = 40$)
2. $\tau_1 < \mu_k^g/\mu_k^r < \tau_2$, too similar ($\tau_1 = 0.9$ and $\tau_2 = 1.1$)
3. $|A - \mu_A| > \sigma_A$, too large or too small areas.

The first inequality selects colorless nuclei and the second nuclei at the transition region. The third condition rejects nuclei with aberrant sizes where μ_A and σ_A are the mean and standard deviation of nuclei areas (A), respectively. The resulting nuclei used to train the classifier are displayed in Fig. 2d and the respective labels were obtained according to Algorithm 1:

Algorithm 1: Automatic algorithm to assign a label to each nucleus based on molecular features.
1: if $\mu_k^g > \mu_k^r$ then label = S/G$_2$
2: if $\mu_k^r > \mu_k^g$ then label = G$_1$

In this work, a total of 3553 nuclei were considered for analysis, with 2291 nuclei (64.5%) labeled as G_1 and 1262 nuclei (35.5%) labeled as S/G_2.

Comparison of cell cycle labels obtained automatically and by visual inspection. In this study, we have developed an automatic procedure to generate cell cycle labels obtained from Fucci2 data, required to subsequently train the classifier. This approach aimed to avoid difficulties associated with interpretation of Fucci2 images by visual inspection. Specifically, these difficulties can occur upon (i) low intensity (dark) nuclei; (ii) transition states (G_1 to S) due to co-expression of cell cycle reporters; or (iii) non-homogeneous cases in which nuclei may exhibit different colors in different nuclear structures. Thus, to overcome this limitation, an automatic strategy was proposed. Herein, the clustering results obtained automatically using Algorithm 1 (A_1) were compared with the labels obtained by visual analysis (VA) based on the Fucci technology. Of note that only nuclei classified by VA were considered for further analysis, which did not include all nuclei classified by A_1. The results indicate that from a total of 2681 nuclei, only 21 nuclei diverged in the classification, corresponding to an accuracy of 99.22%. Specifically, as shown in Figs. 3a, 5 nuclei automatically labeled by A_1 as G_1 were labeled as S/G_2 by VA, and 16 nuclei labeled by the A_1 as S/G_2 were labeled by VA as G_1. The high accuracy obtained is explained by the fact that VA only considered nuclei that could be clearly classified. Figure 3b displays the 21 nuclei labeled differently by A_1 and VA, with the 16 nuclei above the blue line labeled by VA as G_1 and the 5 nuclei below this line labeled by VA as S/G_2. Importantly, the blue line represents the bisectrix of the first quadrant that roughly separates G_1 and G_2 classes. An example of a Fucci2 image and the corresponding labels obtained by VA is shown in Fig. 3c. The divergences between the automatic method of label detection and the ground truth provided by VA are possibly due to misclassifications related to color miss-perception inherent to the human operator that the automatic method solved after color normalization. Overall, these results suggest that the automatic proposed method has higher accuracy than the manual approach using visual inspection and is much less time consuming, thus improving our strategy.

Automatic cell cycle staging and comparison between Fucci2 labels and DAPI features. Our strategy to analyze the interphase cell cycle stages takes advantage of the common DNA binding dye DAPI. In this work, in order to further evaluate nuclei labeled by Algorithm 1, a cell cycle profile was generated from total DAPI intensity, as shown in Fig. 3d. The histogram, which was obtained from the distribution of labeled nuclei within the distinct phases, reveals a higher number of nuclei in G_1 and higher DNA content in S/G_2 than in G_1 (Fig. 3d), which is consistent with a typical cell cycle phase distribution. Overall, these results support the use of our automatic approach in the generation of cell cycle labels.

Moreover, in order to further study the association between Fucci2 and DAPI data, we have represented the 3553 labeled nuclei in the 2D space (normalized area, normalized DAPI intensity). As shown in Fig. 4, the results demonstrate that green nuclei in S/G_2, as determined by Fucci technology, exhibit higher DAPI intensity and larger area than red nuclei in G_1. Taken together, these results corroborate our previous data[21] and strongly support the variations in DNA content along the cell cycle, which correlate with DAPI staining. In summary, our results confirmed that the information obtained from the Fucci technology regarding the cell cycle can also be obtained from DAPI nuclear staining.

Nuclei classification based on DAPI features and performance evaluation. Our results indicate that DAPI area and intensity can be used to evaluate the cell cycle. Therefore, a SVM was trained and tested with DAPI features as input, and the ground truth labels were generated from Fucci2 images.

Normalized features, $\bar{f}_i$ were used to train the SVM classifier. These features are zero mean and unit variance normalized versions of the ones in Eqs. (1) (area) and (2) (DAPI intensity), computed from DAPI image planes, according to:

$$\bar{f}_i = \frac{f_i - \mu_f}{\sigma_f}, \tag{9}$$

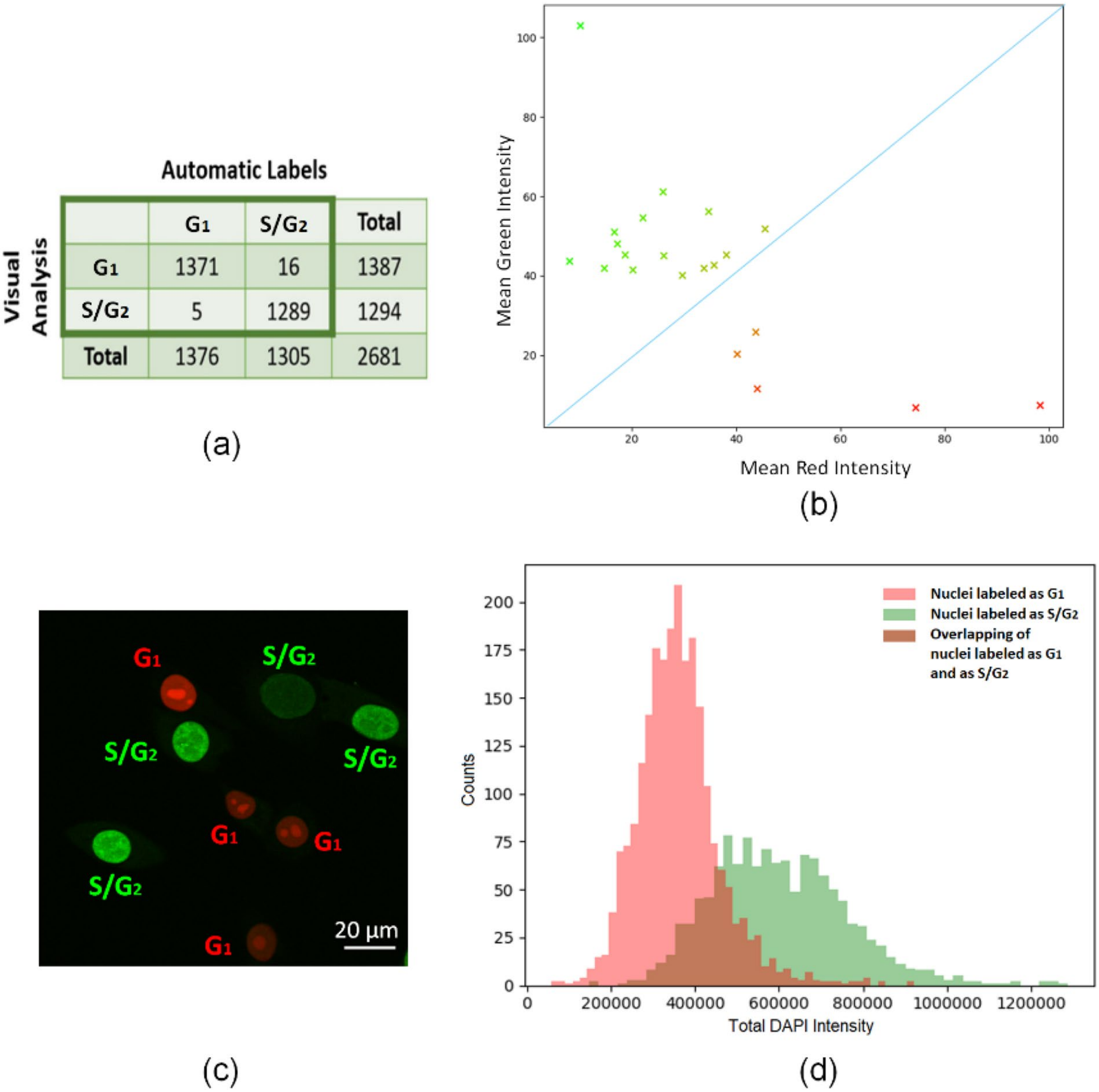
where μ_f and σ_f are the mean and the standard deviation corresponding to feature f_i, computed for each image separately.

Finally, we assessed cell cycle staging on new DAPI images. Firstly, nuclei segmentation was performed and features were extracted from DAPI images (intensity) and from the corresponding segmentation masks (area). Thereafter, these features were normalized according to Eq. (9), by computing μ_f and σ_f for nuclei area and intensity. These normalized features were then fed to the classifier that returned the cell cycle phase (G_1 or S/G_2) for each nucleus.

Nuclei classification was subsequently validated and the estimation of prediction error was performed using both nested five-fold cross-validation and nested leave-one-experiment-out cross-validation.

Nested five-fold cross-validation. The performance of the nuclei classification strategy was evaluated using Precision, Recall and F1-Score, as shown in Table 2. In this table, data represent Precision, Recall and F1-Score when class G_1 or class S/G_2 were considered positive, as well as the average between both classes. The values correspond to the mean and standard deviation over the five models obtained by performing nested five-fold cross-validation.

As shown in Table 2, the proposed approach for cell cycle staging can be performed based on features extracted from DAPI images, as demonstrated by high Precision, Recall and F1-Scores. Specifically, the results



Automatic Labels

Visual Analysis	G1	S/G2	Total
G1	1371	16	1387
S/G2	5	1289	1294
Total	1376	1305	2681

Figure 3. Identification of labels obtained automatically and by visual inspection. **(a)** Confusion matrix between the labels generated according to Algorithm 1 and by visual analysis. **(b)** Representation of each nucleus, labeled by visual analysis, in the 2D space $[\mu_k^r, \mu_k^g]$. Each point is colored with the following intensities $(R, G, B) = \left[\overline{\mu}_k^r, \overline{\mu}_k^g, 0\right]$. The blue line corresponds to the equation mean green intensity = mean red intensity, and is used to identify mislabeled nuclei evaluated by visual analysis. **(c)** Example of a Fucci2 image and the corresponding manual labels obtained by visual analysis. **(d)** Cell cycle profile of total DAPI intensity for labeled nuclei. The red and green bars indicate counts of nuclei labeled as G_1 and S/G_2, respectively. The brown area shows the overlapping bars of nuclei labeled as G_1 and as S/G_2.

demonstrate that the critical features to be considered are relative area and DAPI intensity, since nuclei in S/G_2 phases have their area increased and present higher DAPI intensity when compared to nuclei in G_1 phase.

Importantly, in this study, the features from all nuclei, from all images were used to train the SVM. The SVM was trained and tested in the normalized input space, with features from different images, which resulted in high F1-scores (F1-Scores 0.915 ± 0.006 and 0.839 ± 0.019 for class G_1 and class S/G_2, respectively). This result can be explained by the fact that this SVM takes into account the relative area and intensity for each nucleus. Thus, this procedure is computationally less demanding when compared to other approaches. Indeed, although we perform feature normalization per image, it only takes about 12 s to compute the normalized area and intensity according to Eq. (9) for all the nuclei analyzed. Overall, our data indicate that the SVM presented in this work can take as input nuclei features from all images to perform cell cycle staging providing an accurate label for each nucleus.

Notably, as observed in Fig. 4a, the most challenging issue is to classify nuclei in G_1 and S/G_2 that have overlapping intensities. Thus, to further evaluate our approach, we have analyzed the performance of our method

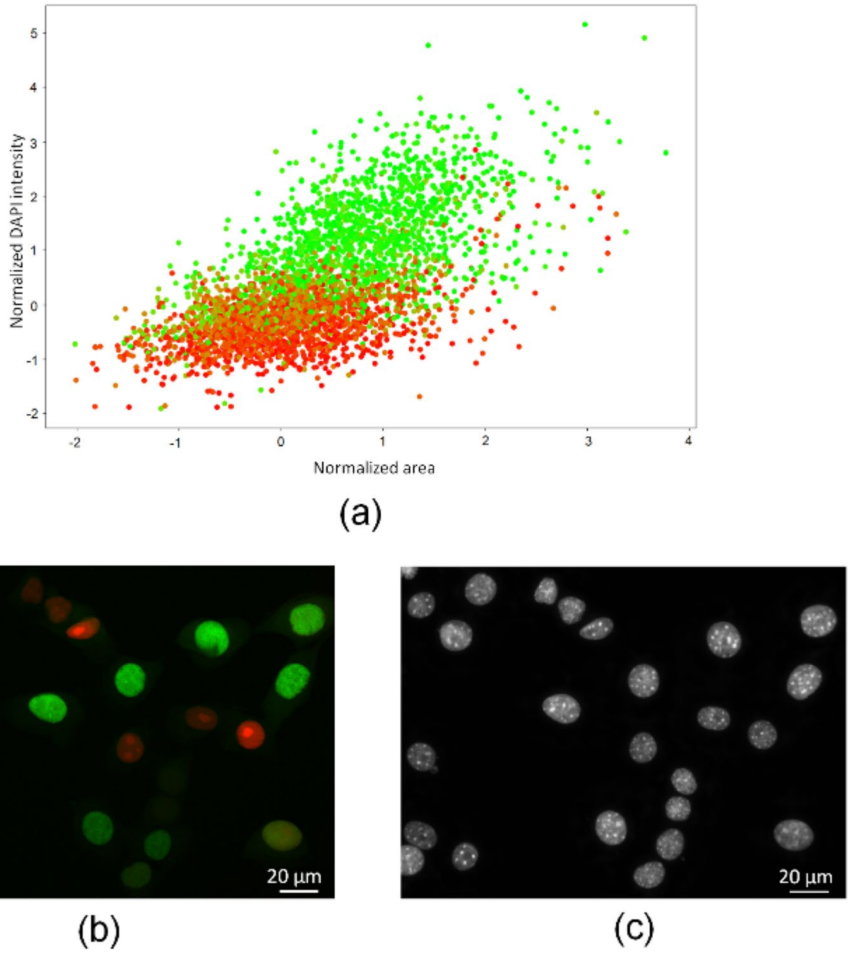


Figure 4. Comparison between Fucci2 labels and DAPI features. Representation of each nucleus in the 2D space (normalized area, normalized DAPI intensity). Each point is colored with the following intensities (R, G, B) = $[\overline{\mu}_k^r, \overline{\mu}_k^g, 0]$, according to the Fucci2 classification. **(b,c)** Representative Fucci2 and DAPI images respectively, illustrating the brightness and size of nuclei in S/G_2 and G_1 stages.

	Precision	Recall	F1-Score
G_1	0.904 ± 0.021	0.926 ± 0.015	0.915 ± 0.006
S/G_2	0.858 ± 0.025	0.822 ± 0.040	0.839 ± 0.019
Average	0.881 ± 0.016	0.874 ± 0.022	0.877 ± 0.010

Table 2. Validation of nuclei classification (μ ± σ). Precision, Recall and F1-Score data are shown for class G_1 positive, class S/G_2 positive and average between both classes. Data represent mean ± standard deviation over the five models obtained by performing nested five-fold cross-validation.

in nuclei with normalized DAPI intensities ranging from -0.5 to 0.5. This subset of data included 525 nuclei, corresponding to approximately 15% of our data set. Precision, Recall and F1-Score obtained for this subset were 0.72, 0.77 and 0.74, respectively. Therefore, even for this challenging subset, our approach was able to achieve an F1-Score of 74%.

Nested leave-one-experiment-out cross-validation. Next, to further assess our strategy, a nested leave-one-experiment-out cross-validation was performed. More specifically, we conducted a nested 13-fold cross-validation to understand how the feature normalization step influences the performance of the proposed approach in each image. A total of 130 images were analyzed and the corresponding 130 values of average F1-Score between class G_1 and S/G_2 are represented in Fig. 5. Remarkably, in 21 of the 130 images, the classifier assigned the correct class to all nuclei (average F1-Score of 100%). Furthermore, as observed in Fig. 5a,b, the distribution of the data is not symmetric. Indeed, in 91 images a high F1-Score was obtained, which was equal or higher than 80%.

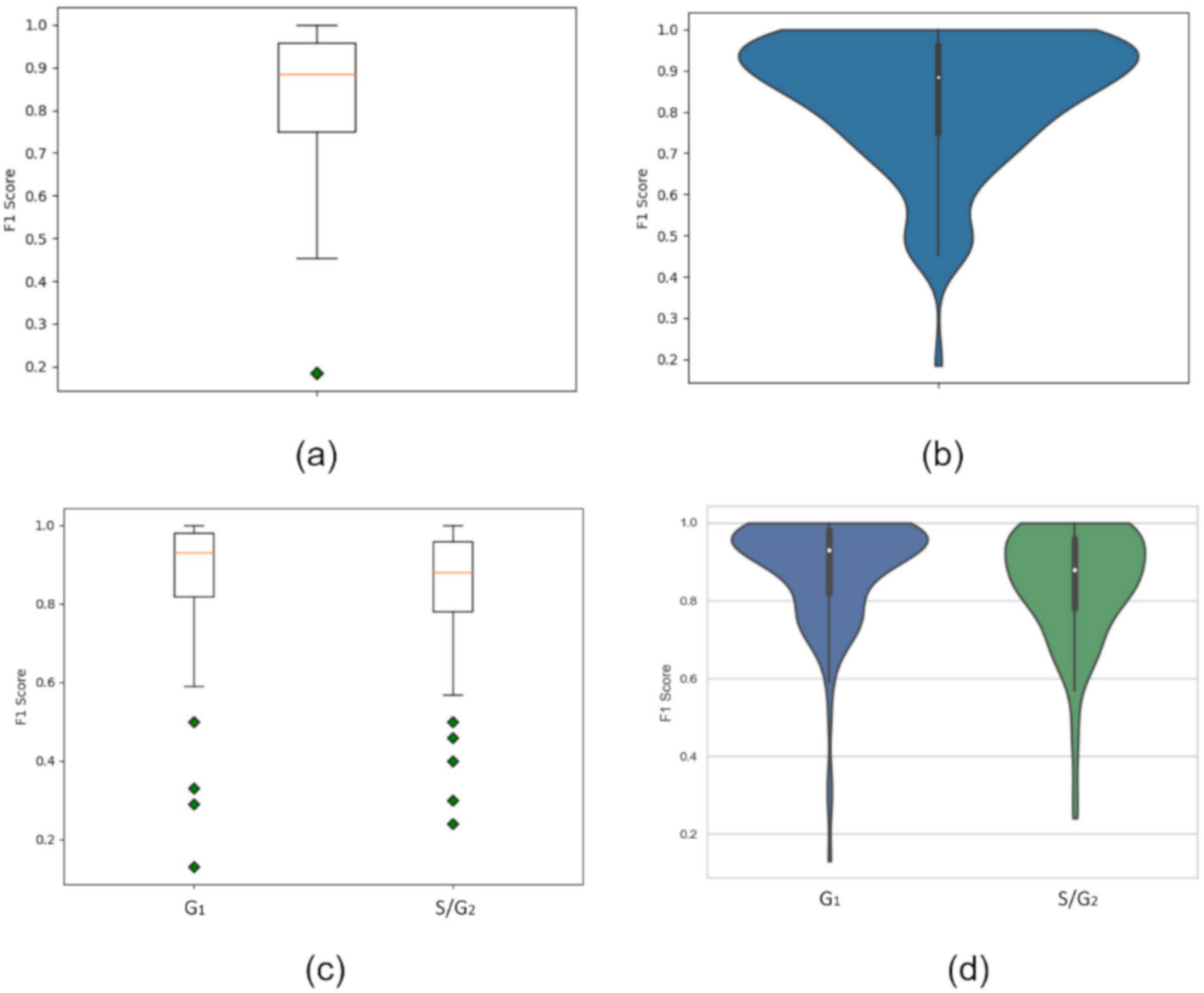


Figure 5. Distribution of nuclei classification **(a)** box plot of F1-Score (130 images). **(b)** Violin plot of F1-Score (130 images). **(c)** Box plot of F1-Score for class G_1 and class S/G_2. **(d)** Violin plot of F1-Score for class G_1 and for class S/G_2.

Moreover, as observed in Fig. 5b, most of the values are clustered around the maximum, that is, the peak with the highest amplitude is located at an F1-Score of approximately 93%. These results show that, for most of the images, our classifier provides proper classification results.

Noteworthy, in this study, feature normalization was performed considering that variations in image acquisition can occur even using the same acquisition parameters. More specifically, area and intensity normalization were conducted for each image and through this normalization step, the relative area and intensity between nuclei in G_1 phase and nuclei in S/G_2 phases were obtained. However, in images with all nuclei at a particular stage, the normalization step could only provide the relative area and intensity of the nuclei at that particular stage, and no information regarding nuclei at the other stage could be made available. Interestingly, the few images with nuclei in one of the classes, corresponded to the images for which an F1-Score lower than 50% was obtained, as observed in Fig. 5b. Moreover, in Fig. 5a, we can detect the presence of an outlier that corresponds to an image with 1 nucleus in G_1 and 15 nuclei in S/G_2. An alternative strategy to avoid this problem, is to perform the normalization per experiment rather than per image, to guarantee enough nuclei in G_1 and in S/G_2. Still, this normalization step may also be affected by acquisition differences between the images.

In addition, box and violin plots of F1-Score for class G_1 and class S/G_2 are shown in Fig. 5b,d. Notably, the data demonstrate better results for class G_1 than those obtained for class S/G_2 (Fig. 5d). This can be explained by the fact that our dataset is imbalanced. As previously described in Fig. 5a, the outliers in Fig. 5b correspond to images that have the majority or all nuclei in one of the two possible classes. By examining the violin plots in Fig. 5d, it can be concluded that for about 75% of the images the F1-Scores for class G_1 are higher than 80% and for about 75% of the images, the F1-Scores for class S/G_2 are higher than 78%. Importantly, for both classes the maximum F1-Score obtained is 100%.

Validation of the classifier on distinct cell types. In order to study the potential of the SVM for cell cycle staging in different cell types, we evaluated a panel of human gastric (AGS, MKN74), breast (MDA-MB-231, MCF7) and colorectal cancer (SW480, HCT116) cell lines, distinct from the cell type used to train it. Specifically,

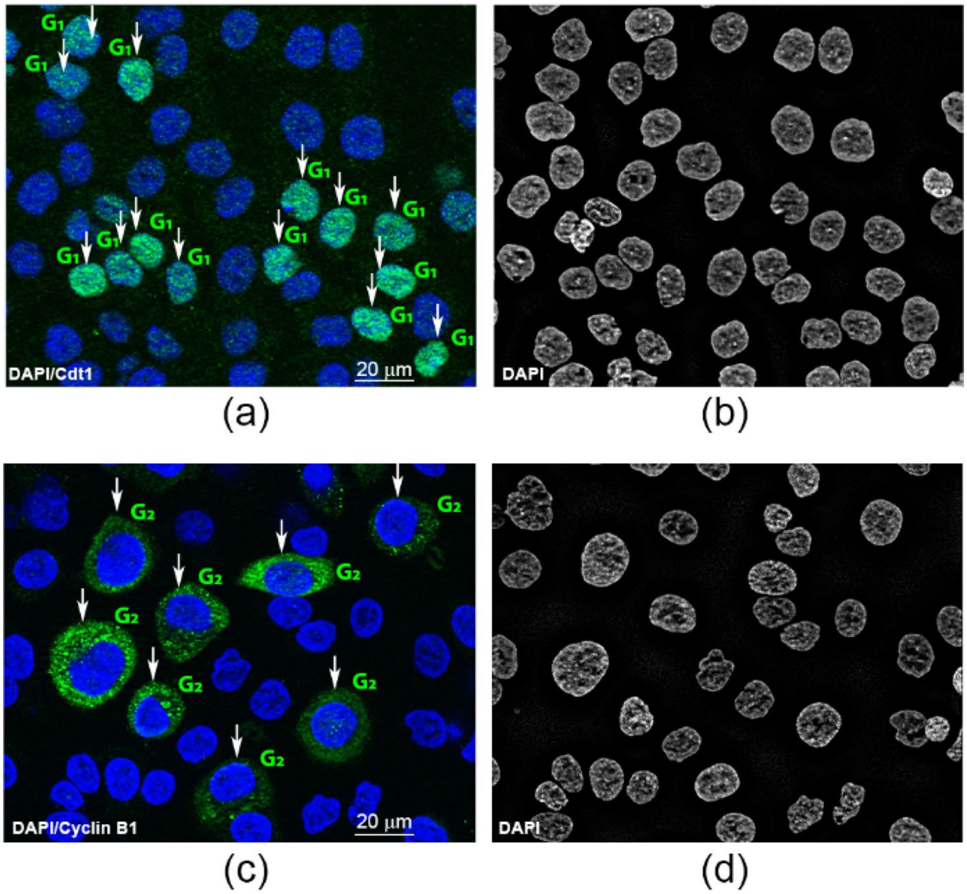


Figure 6. Method applicability in immunofluorescence images of gastric cancer cells. Human gastric cancer cells AGS were stained with Cdt1 (green) **(a)** or cyclin B1 (green) **(c)** and counterstained with DAPI (blue). White arrows indicate cells considered positive for Cdt1 or cyclin B1 expression and used for further analysis and G_1 and G_2 denotes the labels obtained after automatic classification based on DAPI staining. **(b,d)** Corresponding DAPI images used for cell cycle staging.

the SVM was trained with all nuclei (3553 nuclei) and its performance was tested on images of human gastric, breast and colorectal cancer cells. Cells were stained with DAPI for cell cycle staging and with Cdt1 or cyclin B1 for validation purposes. An example is shown in Fig. 6, with Cdt1 positive cells labelled in green and nuclear DAPI in blue (Fig. 6a). Similarly, Fig. 6c illustrates cyclin B1 positive cells labelled in green and nuclear DAPI in blue. The corresponding DAPI images used for cell cycle staging are presented in Fig. 6b and Fig. 6d. Nuclei in G_1 and G_2 were annotated manually based on the expression of Cdt1 or cyclin B1, markers of G_1 and G_2/M phases, respectively. Note that mitotic nuclei were not considered in this study and therefore nuclei likely to be at the M phase were excluded. Moreover, only nuclei with strong Cdt1 and cyclin B1 fluorescence intensity were selected for analysis, as ascertained by visual inspection. This dataset included 180 images, comprising 15 images for each cell type, for each specific antibody. For analysis, nuclei were segmented, features area and DAPI total intensity were computed and normalized, and each segmented nucleus was classified using the SVM previously trained. Interestingly, high recall values (ranging from 89.1 to 99.5%) were obtained for all cell lines analyzed. Specifically, recall values were as follow: AGS (Cdt1: 93.8%; cyclin B1: 93.3%), MKN74 (Cdt1: 99.5%; cyclin B1: 93.2%), MDA-MB-231 (Cdt1: 90.1%; cyclin B1: 89.1%), MCF7 (Cdt1: 89,6%; cyclin B1: 91.8%), SW480 (Cdt1: 94.3%; cyclin B1: 96.1%), HCT116 (Cdt1: 91.1%; cyclin B1: 96.9%). Taken together, the data strongly support that this method can be used for cell cycle staging of distinct cell types.

Discussion

Cell cycle evaluation has long been a prime issue in biological and cancer research and the focus of many clinical studies. However, and despite technical advances, cell cycle staging remains challenging, namely using automated methods in non-manipulated cells and for single cell analysis. Over the years, extensive efforts have been made to improve available approaches but most methods still rely on the evaluation of cell populations and involve the use of cell cycle markers and extensive cell manipulation that disrupt the natural cellular architecture[12]. Importantly, both limitations account for the lack of applications in minimal invasive methodologies in cancer diagnosis and monitoring, as cytology and CTCs analysis. Recently, novel strategies have emerged involving automated imaging analysis in the field of computational life sciences. Distinct frameworks have been proposed in recent years[17,19,22,24–27]. Still, most of the available procedures are not suitable for in situ analysis and interphase cell cycle staging at single cell level.

In this study, we propose a new pipeline for interphase cell cycle staging of single cells that is based on nuclear features using in situ cell images and that can be applied to heterogeneous cell populations. In particular, we developed a new bioimaging pipeline using in situ fluorescence images of nuclei stained solely with DAPI, involving a deep learning segmentation procedure followed by a supervised classifier (SVM) for phase identification. In order to establish the SVM, training was achieved using molecular ground truth labels. Indeed, for training purposes, we have specifically developed an automatic procedure to identify cell cycle labels based on Fucci data to overcome the limitations associated with classification by visual inspection. Our results have shown that the automatic method is more accurate and less time consuming than the manual approach, which are two fundamental properties in transferring new technologies into clinical applications or to be translated into innovation platforms for high throughput drug discovery.

Notably, our strategy takes advantage of the extensively used DNA dye DAPI that exhibits strong fluorescence when bound to adenine–thymine rich sequences of DNA and that correlates with DNA content along the cell cycle[12,28]. Noteworthy, our data has shown a typical cell cycle histogram following automatic cell cycle staging based on DAPI, further supporting our approach. Moreover, DAPI area and intensity predicted cell cycle staging, in a simple and easier strategy, while avoiding the use of cell cycle markers and laboratory manipulation. Remarkably, high Precision, Recall and F1-score were achieved upon performance evaluation. In addition, our classifier was further validated on a panel of six distinct human cancer cells types including gastric, breast and colorectal cancer cells, solely stained with DAPI. Cell cycle regulatory proteins are expressed during particular phases of the cell cycle and are often used to identify specific phases, as Cdt1[13,15] and cyclin B1[11–13,15]. Thus, in order to obtain the ground truths, cells were stained with DAPI and Cdt1 or cyclin B1, specific cell cycle markers that were used to annotate nuclei in G_1 or G_2, respectively. The results demonstrate that the SVM was able to properly classify nuclei in G_1 and G_2 based on DAPI, independently of the cell type. Remarkably, our approach was able to achieve high recall values for all cell lines, despite different nuclear sizes specific to each cell line, highlighting the importance of the feature normalization step proposed in this work. Therefore, although the SVM was trained with features from murine mammary gland cells stained with DAPI, the SVM can also be used for cell cycle staging of other cell types.

When compared to previously described methods, our cell cycle staging pipeline holds advantages. For instance, in the method developed by Roukos et al.[22] the validation step involved comparison of the cell cycle distributions with the ones obtained by flow cytometry[22]. In contrast, in this work, a nucleus by nucleus validation was performed and for each nucleus the predicted G_1 or S/G_2 cell cycle phase was compared with the ground truth label of that nucleus. Moreover, in the protocol proposed by Roukos et al. only one DAPI feature was used, the integrated intensity, which depends on the experimental conditions[22]. Furthermore, although the method allows cell cycle staging of individual cells, it is based on manually defined thresholds, which may be subjective[22]. Hence, compared to the work presented by Roukos et al.[22], the proposed approach has the advantage of using automatic methods for cell cycle staging. In addition, our former framework, also presents some limitations. In that approach, area and total DAPI intensity were calculated for each nucleus, which did not result from automatic nuclei segmentation, and the obtained features formed the feature space in which a modified K-means clustering was performed achieving an overall sensitivity of 94%[21]. However, it relied on features that may vary between different experiments and different cell lines, and thus the K-means algorithm needed to be applied image by image[21]. Thus, our previous method[21] is more time-consuming compared to the proposed approach. Furthermore, in our previous work[21] we did not study the performance of the method on other cell lines and therefore cannot conclude about its generalization capability. Importantly, the major drawback of both the above mentioned methods is the nuclei segmentation step, since in both cases it relies on traditional nuclei segmentation methods which can be sensitive to noise and depend on manual parameter tuning[29]. In contrast, we used a deep learning based segmentation method in which nuclei were first detected through the Fast YOLO architecture[23]. Moreover, in our previous work, the ground truth labels generated to validate the performance of the method resulted from an intensive work of the observer with inevitable pitfalls[21]. In addition to these, other data analysis workflows are becoming available including DeepFlow[26]. However, a major drawback of this approach is the lower accuracy in detecting interphase cells (79.40%) as compared to cells in mitosis (98.73%), which demonstrates the difficulty to distinguish cells at interphase. Furthermore, this technique alters the natural architecture of adherent cells, since it requires cells in suspension[26], and therefore it is not suitable for cell cycle staging of adherent cells in situ. One of the most challenging issues in cell cycle staging is the classification of G_1 and S/G_2 nuclei with overlapping intensities. However, even for this nuclei subset, our approach achieved similar results obtained by others[26]. In the future, additional features may be extracted, such as textural features, to further distinguish this subset of nuclei.

Overall, our data strongly suggest that this method is cell independent, reliable and robust and that in the future it could be used to evaluate single cells within cell populations, which can clearly have an impact in cancer research and clinical applications. Ultimately, we anticipate the proposed pipeline can have major implications in disease monitoring, in the development of novel cell cycle inhibitors in large scale screens, as well as in the selection of appropriate therapeutic strategies along disease progression namely using CTCs analysis.

Materials and methods

This section describes in detail the supervised approach used for interphase cell cycle staging and the datasets required for testing and training the classifier.

Datasets. In this study, two distinct datasets were used to test, train and validate the classifier. The first includes data obtained as previously described by Ferro et al. and comprises a set of 130 fluorescence microscopy images, from a panel of thirteen datasets of images (corresponding to eight independent cell passages), with more than 3500 nuclei with a resolution of 1040 × 1388 pixels[21]. This is the same dataset used by Ferro et al., in which a non-supervised approach for cell-cycle staging was proposed[21]. The images are from murine mammary gland NMuMG-Fucci2 cells (RCB2868, RIKEN Cell Bank, Japan) constitutively expressing Fucci2 probes and stained with DAPI. Briefly, cells cultured on glass coverslips were fixed and stained with DAPI and images were acquired using the acquisition settings described in[21]. These are organized as RGB images with Fucci2 data on the red and green channels and DAPI information on the blue channel. An additional set of 180 images from human gastric (AGS, MKN74), breast (MDA-MB-231, MCF7) and colorectal (SW480, HCT116) cancer cells, stained with DAPI and Cdt1 or cyclin B1 as described in Materials and Methods, was also used to experimentally validate the proposed method. This dataset included 30 images of each cell type, 15 labeled with Cdt1 and 15 with cyclin B1. The use of a completely different set of images as input to the classifier was performed for generalization ability assessment.

Fucci2 image processing and image segmentation. All Fucci2 images were subject to an image processing pipeline in order to automatically compute the labels for the classifier. For background removal, a pre-processing framework was applied to all images. Briefly, for each color channel of each image, the mean background intensity was subtracted. For each image the ground truth nuclei segmentation mask was obtained from manual annotation. Based on the mask for each image, the average background's intensity for each color channel was calculated and denoted as $(\overline{\mathrm{R}}, \overline{\mathrm{G}}, \overline{\mathrm{B}})_{\mathrm{background}}$. Thereafter, every pixel in the image was obtained by subtracting to its intensity $((\mathrm{R,G,B})_{\mathrm{pixel}})$ the average background's intensity per channel. That is, by computing the following quantity: $(\mathrm{R,G,B})_{\mathrm{pixel}} - (\overline{\mathrm{R}}, \overline{\mathrm{G}}, \overline{\mathrm{B}})_{\mathrm{background}}$. Subsequently, nuclei segmentation was performed for all images. Specifically, we have used our recently developed deep learning based segmentation method, in which nuclei are first detected in the images using the Fast YOLO architecture[23]. Afterwards, patches corresponding to the detected nuclei were extracted from the image, resized to a fixed size and used as input of a U-Net, which will compute the corresponding segmentation mask of the patch that is resized back to its original size[23]. Notably, nuclei at the borders were not considered for analysis, since the provided information may be incomplete. The segmentation experiments were carried out on a NVIDIA GPU GTX 1050 (4 GB) and Python 3.6. The deep learning implementations were based on the open-source deep learning libraries Tensorflow and Keras. The segmentation model described in this section is available at https://github.com/HemaxiN/YOLO_UNET. Training of the deep learning approach for Fucci2 images was performed as described in our previous work[23]. For these images, a F1-Score above 0.8 was obtained for IoU thresholds below 0.75. To segment the DAPI images, the model trained in[23] was used.

Nuclei classification from DAPI features. For nuclei classification, normalized DAPI features were used to train the SVM classifier[30]. Thus we trained and evaluated the performance of the SVM in the 2D input space of normalized area and normalized DAPI intensity. Hyperparameter optimization was performed in the following parameter space:

- Kernel: 'rbf'; Gamma: (1e−2, 1e−3, 1e−4, 1e−5); C: (0.001, 0.10, 0.1, 10, 25, 50, 100, 1000);
- Kernel: 'poly'; C: (0.001, 0.10, 0.1, 10, 25, 50, 100, 1000); degree: (1, 2, 3, 4, 5);
- Kernel: 'sigmoid'; Gamma: (1e−2, 1e−3, 1e−4, 1e−5); C: (0.001, 0.10, 0.1, 10, 25, 50, 100, 1000);
- Kernel: 'linear'; C: (0.001, 0.10, 0.1, 10, 25, 50, 100, 1000);

To perform the nested five-fold cross-validation the entire dataset was divided into five folds (A, B, C, D and E). The training and testing of the model was performed five times, considering each of the folds (A, B, C, D and E) as test set once. A schematic representation of the five-fold cross-validation is shown in Supplementary Fig. S2. For instance, considering fold E as test set, a model was trained using folds A, B, C and D. The training/validation split is 80%/20%, and for each parameter combination, a model was built using the training set and its performance was evaluated on the validation set. The best model is the one that presents the highest score (average F1-Score between G_1 and S/G_2) on the validation set. Finally, the performance of this model was evaluated on the test set (fold E in this example). This procedure was repeated five times (considering folds A, B, C, D and E as test set at a time), and the final results represent the average ± standard deviation of the five test folds.

Furthermore, we also performed a nested 13-fold cross-validation since the dataset includes images from 13 experiments, equivalent to a nested leave-one-experiment-out cross-validation. That is, the model was trained with nuclei from 12 experiments, and its performance was tested on nuclei from the other experiment. This process was repeated 13 times. The average F1-Score values between class G_1 and S/G_2 were represented in box and violin plots in order to understand the distribution of the data. As stated before, the features of the proposed input space were normalized image by image.

Additionally, for both experiments, class weights were set inversely proportional to the class frequencies in the training data. These weights are inversely proportional to the class frequencies in the training data:

$$cweight\_i = \frac{nsamples}{nclassesxnsamples\_i} \tag{10}$$

where cweight_i denotes the class weight for the class i, nsamples the number of samples in the training data, nclasses the number of classes in the classification problem, and nsamples_i the number of samples belonging to class i. This is a technique used to deal with imbalanced datasets, and it will train the SVM in training samples with different cost weights in the objective function[31]. Herein, there are more nuclei in G_1 phase than in S/G_2 phases, and therefore the class weight is higher for S/G_2 nuclei. The classification experiments were carried out in Python 3.6 using the SVM implementation available in scikit learn (version 0.21.2). The cell cycle staging algorithm developed in this work is available at https://github.com/HemaxiN/InterphaseCellCycleStaging.

Cell culture, immunofluorescence staining and image acquisition. Human gastric (AGS, MKN74), breast (MDA-MB-231, MCF7) and colorectal (SW480, HCT116) cancer cells were obtained from the American Type Culture Collection (ATCC). Briefly, cells were grown in RPMI 1640 (AGS, MKN74, SW480, HCT116), DMEM (MDA-MB-231) or DMEM/F-12 (MCF7) (Gibco, Invitrogen) supplemented with 10% fetal bovine serum (Hyclone) and 1% penicillin/ streptomycin (Gibco, Invitrogen) in a humidified incubator at 37 °C, 5% CO_2. Cells were cultured on glass coverslips and fixed with 4% paraformaldehyde for 20 min. Following a 10 min wash in phosphate buffered saline (PBS), cells were permeabilized with 0.1% Triton X-100 in PBS for 15 min at room temperature. Cells were blocked with 3% bovine serum albumin (BSA) in PBS and stained overnight at 4 °C with Cdt1 rabbit primary antibody (1:200, Cell Signaling, #8064) or cyclin B1 rabbit primary antibody (1:800, Cell Signaling #12231). Subsequently, cells were incubated with Alexa Fluor 488 goat anti-rabbit (1:200, Invitrogen, Thermo Fisher Scientific) for 1 h in the dark. Nuclei were stained with DAPI (Sigma-Aldrich, 0.1 μg/ml in PBS) for 15 min and coverslips were mounted on slides using Vectashield (Vector Laboratories). Images were acquired on a Carl Zeiss Apotome Axiovert 200 M Fluorescence Microscope (Carl Zeiss, Jena, Germany) with a 40× objective (Plan-Apochromat 40x/1.3 Oil DIC (UV) VIS-IR M27) using an Axiocam HRm camera and the Zeiss Axion Vision 4.8 software. For DAPI staining, multiple images were acquired along the z axis (60 z stacks; 5 ms exposure) and images were deconvoluted using deconvolution express in Huygens Software (Scientific Volume Imaging). All images were acquired with the same acquisition settings. For Cdt1 and cyclin B1 staining, a single plane image was acquired to confirm cells in the G_1 or G_2 phase of the cell cycle, respectively. Image J software was used for analysis.

Received: 19 March 2021; Accepted: 9 September 2021
Published online: 29 September 2021

References

1. Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nat. Rev. Cancer* **17**, 93–115. https://doi.org/10.1038/nrc.2016.138 (2017).
2. Nurse, P. A long twentieth century of the cell cycle and beyond. *Cell* **100**, 71–78. https://doi.org/10.1016/s0092-8674(00)81684-0 (2000).
3. Norbury, C. & Nurse, P. Animal cell cycles and their control. *Annu. Rev. Biochem.* **61**, 441–470. https://doi.org/10.1146/annurev.bi.61.070192.002301 (1992).
4. Vermeulen, K., Van Bockstaele, D. R. & Berneman, Z. N. The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer. *Cell Prolif.* **36**, 131–149. https://doi.org/10.1046/j.1365-2184.2003.00266.x (2003).
5. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: A changing paradigm. *Nat. Rev. Cancer* **9**, 153–166. https://doi.org/10.1038/nrc2602 (2009).
6. Loddo, M. *et al.* Cell-cycle-phase progression analysis identifies unique phenotypes of major prognostic and predictive significance in breast cancer. *Br. J. Cancer* **100**, 959–970. https://doi.org/10.1038/sj.bjc.6604924 (2009).
7. Sommariva, S., Tarricone, R., Lazzeri, M., Ricciardi, W. & Montorsi, F. Prognostic value of the cell cycle progression score in patients with prostate cancer: A systematic review and meta-analysis. *Eur. Urol.* **69**, 107–115. https://doi.org/10.1016/j.eururo.2014.11.038 (2016).
8. Begnami, M. D., Fregnani, J. H., Nonogaki, S. & Soares, F. A. Evaluation of cell cycle protein expression in gastric cancer: Cyclin B1 expression and its prognostic implication. *Hum. Pathol.* **41**, 1120–1127. https://doi.org/10.1016/j.humpath.2010.01.007 (2010).
9. Dokumcu, K. & Farahani, R. M. Evolution of resistance in cancer: A cell cycle perspective. *Front. Oncol.* **9**, 376. https://doi.org/10.3389/fonc.2019.00376 (2019).
10. Hallett, R. M. *et al.* Treatment-induced cell cycle kinetics dictate tumor response to chemotherapy. *Oncotarget* **6**, 7040–7052. https://doi.org/10.18632/oncotarget.3140 (2015).
11. Sherr, C. J. & Bartek, J. Cell cycle-targeted cancer therapies. *Annu. Rev. Cancer Biol.* **1**, 41–57 (2017).
12. Eastman, A. E. & Guo, S. The palette of techniques for cell cycle analysis. *FEBS Lett.* https://doi.org/10.1002/1873-3468.13842 (2020).
13. Sakaue-Sawano, A. *et al.* Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498. https://doi.org/10.1016/j.cell.2007.12.033 (2008).
14. Sakaue-Sawano, A., Kobayashi, T., Ohtawa, K. & Miyawaki, A. Drug-induced cell cycle modulation leading to cell-cycle arrest, nuclear mis-segregation, or endoreplication. *BMC Cell Biol.* **12**, 2. https://doi.org/10.1186/1471-2121-12-2 (2011).
15. Sakaue-Sawano, A. & Miyawaki, A. Visualizing spatiotemporal dynamics of multicellular cell-cycle progressions with fucci technology. *Cold Spring Harb. Protoc.* https://doi.org/10.1101/pdb.prot080408 (2014).
16. Jevtic, P., Edens, L. J., Vukovic, L. D. & Levy, D. L. Sizing and shaping the nucleus: Mechanisms and significance. *Curr. Opin. Cell Biol.* **28**, 16–27. https://doi.org/10.1016/j.ceb.2014.01.003 (2014).
17. Blasi, T. *et al.* Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nat. Commun.* **7**, 10256. https://doi.org/10.1038/ncomms10256 (2016).
18. Chen, X., Zhou, X. & Wong, S. T. Automated segmentation, classification, and tracking of cancer cell nuclei in time-lapse microscopy. *IEEE Trans. Biomed. Eng.* **53**, 762–766. https://doi.org/10.1109/TBME.2006.870201 (2006).
19. Wang, M. *et al.* Novel cell segmentation and online SVM for cell cycle phase identification in automated microscopy. *Bioinformatics* **24**, 94–101. https://doi.org/10.1093/bioinformatics/btm530 (2008).
20. Yan, J. *et al.* An effective system for optical microscopy cell image segmentation, tracking and cell phase identification. In *IEEE International Conference on Image Processing, ICIP; 1917–1920.* https://doi.org/10.1109/ICIP.2006.313143 (2006).
21. Ferro, A. *et al.* Blue intensity matters for cell cycle profiling in fluorescence DAPI-stained images. *Lab. Invest.* **97**, 615–625. https://doi.org/10.1038/labinvest.2017.13 (2017).

22. Roukos, V., Pegoraro, G., Voss, T. C. & Misteli, T. Cell cycle staging of individual cells by fluorescence microscopy. *Nat. Protoc.* **10**, 334–348. https://doi.org/10.1038/nprot.2015.016 (2015).
23. Narotamo, H., Sanches, J. M. & Silveira, M. Segmentation of cell nuclei in fluorescence microscopy images using deep learning. in *Pattern Recognition and Image Analysis. IbPRIA 2019. Lecture Notes in Computer Science.* Vol. 11867. 53–64. (Morales A., Fierrez J., Sánchez J., Ribeiro B. eds). https://doi.org/10.1007/978-3-030-31332-6_5 (Springer, 2019).
24. Mao, Y., Han, L. & Yin, Z. Cell mitosis event analysis in phase contrast microscopy images using deep learning. *Med. Image Anal.* **57**, 32–43 (2019).
25. Li, F., Zhou, X., Ma, J. & Wong, S. T. Multiple nuclei tracking using integer programming for quantitative cancer cell cycle analysis. *IEEE Trans. Med. Imaging* **29**, 96–105. https://doi.org/10.1109/TMI.2009.2027813 (2010).
26. Eulenberg, P. *et al.* Reconstructing cell cycle and disease progression using deep learning. *Nat. Commun.* **8**, 463. https://doi.org/10.1038/s41467-017-00623-3 (2017).
27. Gomes, C. J., Harman, M. W., Centuori, S. M., Wolgemuth, C. W. & Martinez, J. D. Measuring DNA content in live cells by fluorescence microscopy. *Cell Div.* **13**, 6 (2018).
28. Kapuscinski, J. DAPI: A DNA-specific fluorescent probe. *Biotech. Histochem.* **70**, 220–233. https://doi.org/10.3109/10520299509108199 (1995).
29. Xing, F. & Yang, L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Rev. Biomed. Eng.* **9**, 234–263. https://doi.org/10.1109/RBME.2016.2515127 (2016).
30. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
31. Y.-M., H. & S.-X., D. Weighted support vector machine for classification with uneven training class sizes. *Int. Conf. Mach. Learn. Cybern. (Guangzhou, China)* **7**, 4365–4369. https://doi.org/10.1109/ICMLC.2005.1527706 (2005).

Acknowledgements

This work was supported by FEDER funds through the Operational Programme for Competitiveness Factors (COMPETE 2020), Programa Operacional de Competitividade e Internacionalização (POCI), Programa Operacional Regional do Norte (Norte 2020) and by National Funds through the Portuguese Foundation for Science and Technology (FCT), under the projects PTDC/BBB-IMG/0283/2014, PTDC/BTM-SAL/30383/2017, LARSyS-UIDB/50009/2020, LARSyS-UID/EEA/50009/2019, NORTE-01-0145-FEDER-000029 and doctoral grant SFRH/BD/114687/2016. The authors acknowledge the American Association of Patients with Hereditary Gastric Cancer "No Stomach for Cancer" for funding Seruca's research and the support of the i3S Scientific Platform Advanced Light Microscopy, member of the PPBI (PPBI-POCI-01-0145-FEDER-022122).

Author contributions

H.N. and M.S.F. were responsible for the conception of the experimental system, data acquisition, data analysis and interpretation, and wrote the manuscript. A.M.M and S.M. were involved in laboratory experiments. H.N, M.S. and J.M.S. were responsible for the development of the algorithm and software and data analysis and interpretation. J.M.S. and R.S. were responsible for study conceptualization and design, data interpretation and review of the manuscript. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-98489-5.

Correspondence and requests for materials should be addressed to R.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.