*CrossRef DOI of original article:*

# Histological Grading of Breast Cancer Malignancy using Automated Image Analysis and Subsequent Machine Learning

Dominik Lenz[1], Paulo César Ribeiro Boasquevisque[2], Robson Dettmann Jarske[3], Célio Siman Mafra Nunes[4], Isabela Passos Pereira Quintaes,[5] and Samuel Santana Sodré[6]

[1] Universidade Vila Velha, Espirito Santo, Brazil

*Received: 1 January 1970 Accepted: 1 January 1970 Published: 1 January 1970*

## Abstract

The objective of this study was to determine the histological degree of breast cancer malignancy using the automated principle of machine learning with the free access computer programs CellProfiler and Tanagra.Methods and results: Digital photographs of neoplastic tissue histological slides were obtained from 224 women with breast cancer. The digitized images were transferred to the CellProfiler software and treated according to a predetermined algorithm, resulting in a database exported to the Tanagra software for further automated classification of the histological degree of malignancy. The Kappa index of agreement between the medical pathologist and the automated analysis performed in the Tanagra software was 0.91 for the tubular score, 0.55 for the nuclear score, and 0.49 for the mitotic index score.

***Index terms***— breast cancer; image analysis; machine learning; cellular diagnosis; histological malignancy grade.

## 1 Introduction

ollowing non-melanoma skin cancer, breast cancer is the most common type of cancer among women and the second worldwide, corresponding to 25.2% of all cancers in world statistics and 29.5% in Brazil. Breast cancer is rare in men, representing less than 1% of cases (American cancer society (2019), Instituto Nacional de Cancer, Brazil, 2017).

To successfully treat and control breast cancer in the female population, it is essential to identify risk factors for the disease. Moreover, early diagnosis and immediate access to treatment are decisive conditions for the disease prognosis (American Cancer Society (2019), Instituto Nacional de Cancer, Brazil, 2017).

The histological grade of malignancy proposed by Scarff, Bloom, and Richardson and further modified by Elston and Ellis, known as the Nottingham Classification System, is considered one of the main factors for determining the prognosis of breast cancer **??**Beck et al., 2011, Chen et Machine learning is advantageous due to its potential to gather a large volume of information, once the appropriate accuracy and precision are achieved, on a specific disease in a single digital tool; suppressing the subjectivity of human evaluation with agility in the analysis of the material to be studied, aiming at safe and quick diagnoses, which could even be used as a "second specialized opinion" in cases of greater complexity (Wernick et al., 2010, Mulrane et al., 2008 **??** Jones et al., 2009, Misselwitz et al., 2010).

The present study aimed to perform an automated and reproducible classification of the parameters used by pathologists to diagnose breast cancer: nuclear score, tubular score, and mitotic index. The software used for image analysis and classification (CellProfiler and Tanagra) used for the present study are free. The results obtained by the automated analysis were compared with a pathologist diagnosis **??**Jones et al., 2009, Carpenter et al., 2006, Lamprecht et al., 2007, Lenz et al., 2017).

# 2 II.

# 3 Materials e Methods

# 4 a) The samples-Inclusion and exclusion criteria

The study targeted women with breast cancer and presenting the most frequent histological types: infiltrating ductal carcinoma, invasive lobular carcinoma, and the mixed infiltrating lobular ductal form; who underwent surgical treatment for this disease in 2015 and that, until the time of surgery, had not undergone adjuvant chemotherapy or radiotherapy treatments. Complete epidemiological diagnosis and treatment data could be obtained, and histological slides were stained by the Hematoxylin & Eosin method with preserved staining, which enabled digital photographs of adequate quality.

The Santa Rita de Cássia Hospital, located in the city of Vitória, is considered the main reference hospital for cancer treatment in the Espírito Santo state, providing medical care for 625 women with breast cancer in 2015.

Out of 276 cases selected for meeting the inclusion and exclusion criteria, 52 patients were also excluded by the pathologist at the Hospital Santa Rita de Cássia due to "in situ" suffering from breast cancers. Since these issues could compromise machine learning and, consequently, the automated analysis of these images, this study included 224 cases at the end.

The year 2015 was selected because the Tumor Record Sheets for that year represents, at the beginning of the study, the most recent and complete data released by the Health Information System -Hospital Cancer Registry of the Ministry of Health of the Federal Government of Brazil.

The

# 5 b) Digitization of histological slides

All histological slides from the 224 selected cases were randomly reviewed by a pathologist without access to patient data at the Hospital Santa Rita de Cássia, aiming to select the samples with the bestpreserved color aspect. Twenty images of breast tissue of each selected patient were obtained using a digital camera (Moticam 1000 1.3 MPixel MTC 1000) attached to a light microscope.

# 6 c) Loading images to CellProfiler

Out of 4,480 digitalized photographs in the 40fold magnification, after their upload to the CellProfiler program, only the artifact-free images were maintained and recognized as adequate by this image analysis program., Therefore, 1937 images were transferred to the CellProfiler software and submitted to its algorithm, These attributes are aspects and characteristics, identified by the CellProfiler software that express the averages of the quantitative parameters of the study's objects (the images) and enabled the automated identification and classification of each object.

# 7 d) CellProfiler algorithm

Following an algorithm developed for treating digitized images for the CellProfiler computational environment, all 1997 images were treated in the following sequence of the 9-step algorithm, as shown in Chart 1.

Chart 1: CellProfiler algorithm.

The 1937 digitized photographs treated according to this algorithm resulted in a data set exported to Tanagra cellular image data analysis software. Then, this dataset was distributed in an Excel spreadsheet (Microsoft R ), and the automated classifications of the tubular, nuclear and mitotic indexes, as well as the histological degree of malignancy, were acquired.

# 8 e) CellProfiler Algorithm i. Phase 1 -Load Images

All the digitized images observed from histological slides at 40-fold magnification were transferred to the CellProfiler software (Figure 1a).

# 9 ii. Phase 2 -Color to Gray

The original scanned images were converted to the white/gray/black spectrum (Figure 1b).

# 10 iii. Phase 3 -ImageMath

Since the CellProfiler software analyzes the study's objects according to the light intensity and the cell nuclei, it was necessary to reverse the nuclei coloration initially stained in black to white and invert the other elements coloring to black (Figure ??c). generated for each digitized image with 47 quantitative parameters, called attributes.

# 11 iv. Phase 4 -Apply Threshold

In this stage, a binary image (i.e., an image with only two-pixel intensities, 0 and 1), was created.

## 12 v. Phase 5 -Identify Primary Objects

Cell nuclei were defined and identified as primary objects of the study in this step of the algorithm (Figure ??d).

## 13 vi. Phase 6 -Measure Objects Size and Shape

Primary objects were measured in this step, and the parameters (attributes), identified by the CellProfiler software for each study object, were acquired by the average of these measurements.

## 14 vii. Phase 7 -Filter Objects

An image filtering was used to suppress changes that could interfere in the primary object analysis, eliminating the artifacts and preserving only the cell nuclei (Figure ??e). After applying the image filter and eliminating artifactual changes, a new measurement of the primary objects (cell nuclei) attributes was performed.

## 15 ix. Phase 9 -Export to Database

After the CellProfiler algorithm steps, 47 quantitative data (attributes) for each primary object studied were identified using qualitative data from the digitized images and defined as parameters, enabling both individual identification and analysis of each primary object.

This list of attributes constituted the database exported to the Tanagra image data analysis software.

## 16 f) Classification after machine learning

Tanagra is open-source software for database analysis and statistical analysis developed under the design of machine learning.

In the present study, Tanagra software was used to perform the automated classification of the malignancy degree of breast cancers for the tubular, nuclear and mitotic index scores, as well as for the histological grade. Moreover, 3 parameters used in the definition of the histological grade in breast cancer were analyzed: the tubular aspect, the nuclear morphology, and the cell count in mitosis; from the analysis of the database containing 47 quantitative parameters for each analyzed object of the study.

## 17 III.

## 18 Statistical Analysis

The tubular, nuclear, and mitotic index scores, which together define the histological degree of malignancy in breast cancer, were determined. The statistical parameters of Predictive Values, Accuracy, Error, and the Kappa Index of agreement between the pathologist and the medical program analyzer, were also used in this phase. The programs Tanagra and Med Calc were used for statistical processing. The statistical parameters gathered were used to determine the histological degree of malignancy.

IV.

## 19 Results

The present study aimed to perform an automated and reproducible classification of the pathological parameters used to diagnose breast cancer: nuclear score, tubular score, and mitotic index.

The automated classification results are depicted in Table ??, while the outcomes comparing the pathological and the automated diagnoses are shown in Table ??. A scatter plot of the automated classification resulted from machine learning is exhibited in Figure ??.

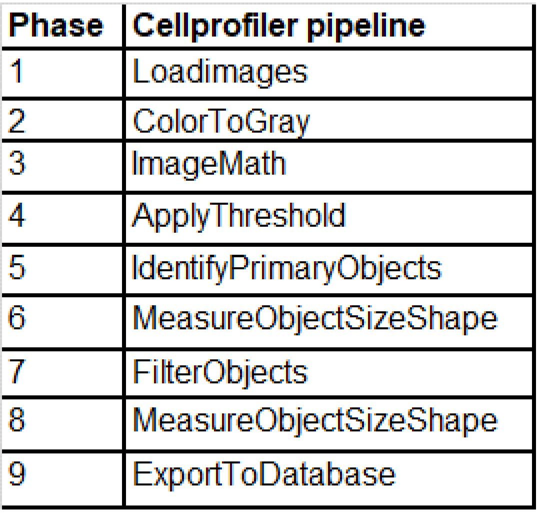
## 20 Discussion

Artificial Intelligence, particularly linked to machine learning, has been increasingly used as a safe and effective tool in disease diagnosis and prognosis, especially on studies assessing breast cancer, a disease of high impact on several women's lives.

This study stands out as a pioneering publication using free access software to diagnose the histological degree of malignancy in breast cancer. Thus, the automated analysis to obtain safe diagnoses of histopathological parameters is a feasible tool since a dataset with sufficient information for adequate machine learning can provide an efficient analysis that ensures remarkable accuracy.

In conclusion, digitalized images of breast cancer histological slides enabled the automated analysis of histopathological parameters, converting them into quantitative parameters for the diagnosis, and defining the histological degree of malignancy. A database expansion is necessary to optimize the analysis and provide the machine sufficient information and data, postulating solid concepts and knowledge to support all requested aspects of the analysis.

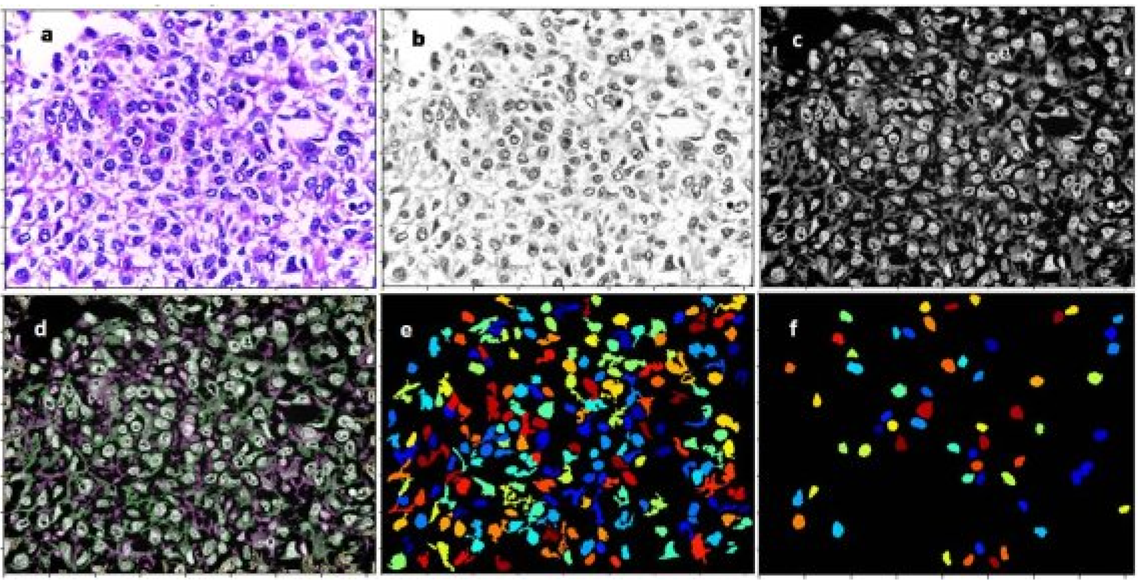
In this sense, further multidisciplinary studies covering machine learning and breast cancer in women may lead to significant novel contributions.



| Phase | Cellprofiler pipeline |
|---|---|
| 1 | Loadimages |
| 2 | ColorToGray |
| 3 | ImageMath |
| 4 | ApplyThreshold |
| 5 | IdentifyPrimaryObjects |
| 6 | MeasureObjectSizeShape |
| 7 | FilterObjects |
| 8 | MeasureObjectSizeShape |
| 9 | ExportToDatabase |

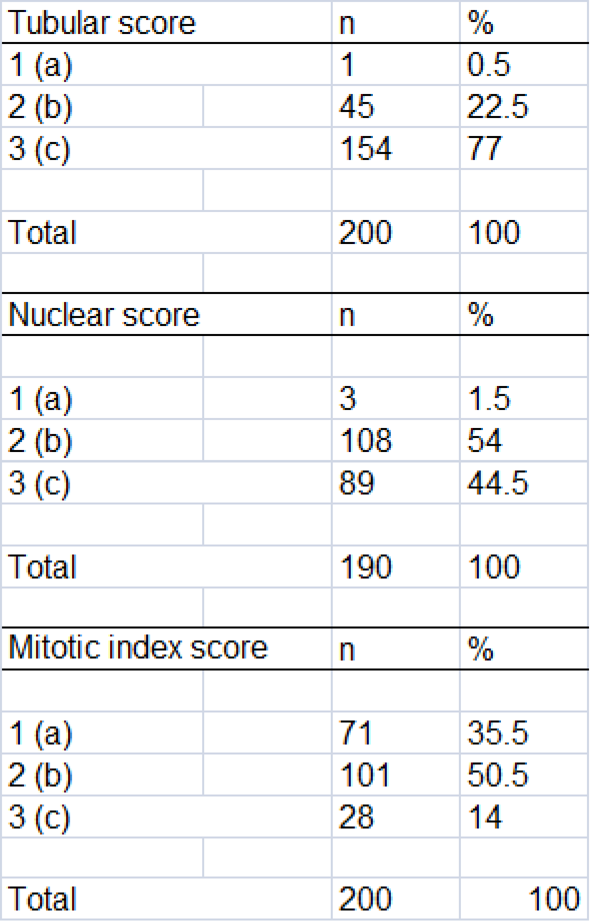
1a

Figure 1: Figure 1a :



1b

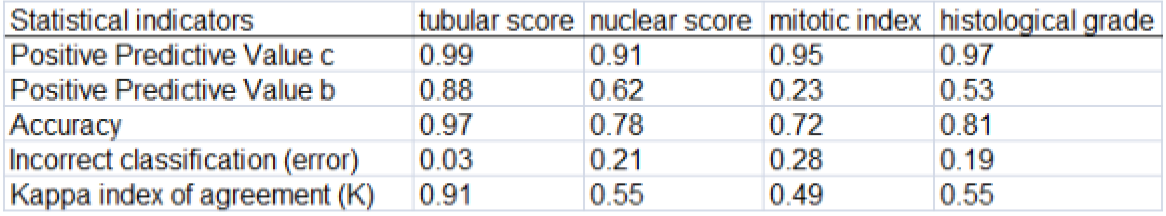
Figure 2: Figure 1b :



| Tubular score | n | % |
|---|---|---|
| 1 (a) | 1 | 0.5 |
| 2 (b) | 45 | 22.5 |
| 3 (c) | 154 | 77 |
| | | |
| Total | 200 | 100 |
| | | |
| Nuclear score | n | % |
| | | |
| 1 (a) | 3 | 1.5 |
| 2 (b) | 108 | 54 |
| 3 (c) | 89 | 44.5 |
| | | |
| Total | 190 | 100 |
| | | |
| Mitotic index score | n | % |
| | | |
| 1 (a) | 71 | 35.5 |
| 2 (b) | 101 | 50.5 |
| 3 (c) | 28 | 14 |
| | | |
| Total | 200 | 100 |

**122a**

Figure 3: Table 1 :Table 2 :Figure 2a :



| Statistical indicators | tubular score | nuclear score | mitotic index | histological grade |
|---|---|---|---|---|
| Positive Predictive Value c | 0.99 | 0.91 | 0.95 | 0.97 |
| Positive Predictive Value b | 0.88 | 0.62 | 0.23 | 0.53 |
| Accuracy | 0.97 | 0.78 | 0.72 | 0.81 |
| Incorrect classification (error) | 0.03 | 0.21 | 0.28 | 0.19 |
| Kappa index of agreement (K) | 0.91 | 0.55 | 0.49 | 0.55 |

Figure 4:

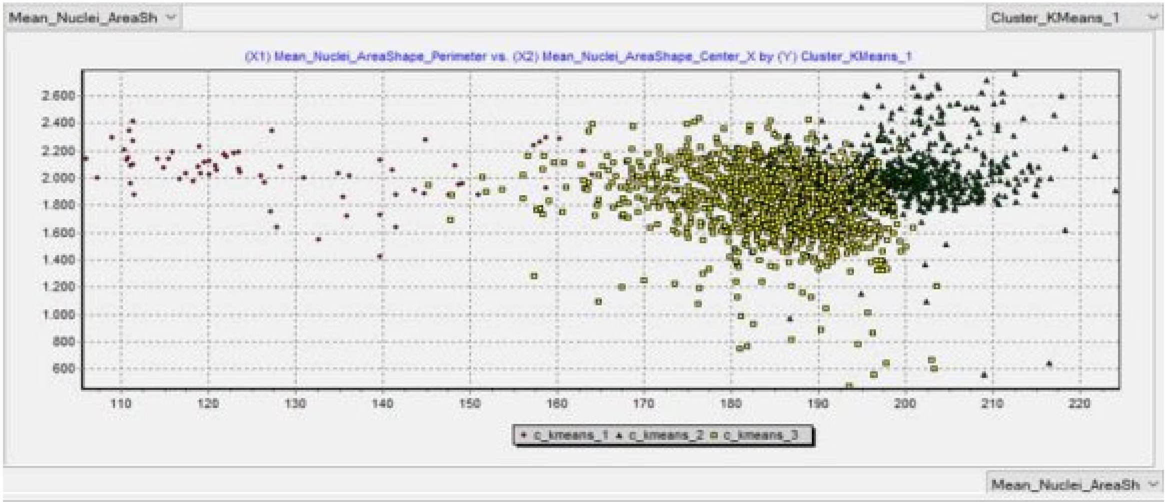


Figure 5:

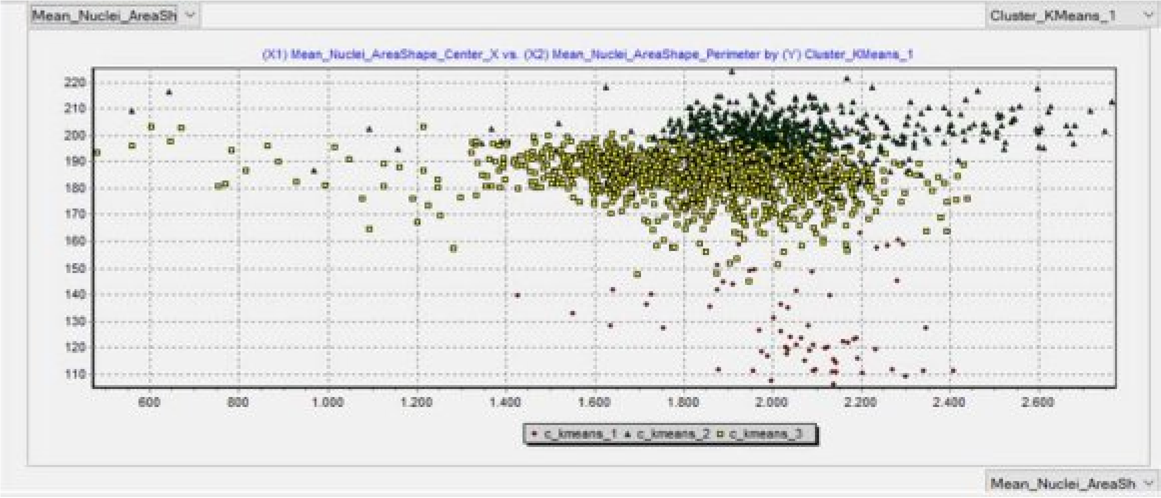


Figure 6:

Figure 7:

Figure 8:

## .1 Conflicts of interest: None declared.

Author contributions: PCRB: Taking images, writing, cooperation with the pathology. RDJ: pathological diagnosis. CSMN: Image analysis, writing. IPPQ: Image analysis, writing. SSS: Image analysis, writing. DL: Supervision, statistical processing, machine learning.

[Procnatlacadsci (2009 10)] , U Procnatlacadsci , SA . 2009 10. 106 p. .

[Romo-Bucheli et al. (2017)] 'A Deep Learning Based Strategy for Identifying and Associating Mitotic Activity with Gene Expression Derived Risk Categories in Estrogen Receptor Positive Breast Cancers'. D Romo-Bucheli , A Janowczyk , H Gilmore , E Romero , A Madabhushi . *Cytometry A* 2017 Jun. 91 (6) p. .

[Dordea et al. ()] 'An open-source computational tool to automatically quantify immunolabeled retinal ganglion cells'. A C Dordea , M A Bray , Allen K Logan , D J Fei , F Malhotra , R Gregory , M S Carpenter , A E Buys , ES . *Exp Eye Res* 2016. 147 p. .

[Hennig et al. ()] 'An open-source solution for advanced imaging flow cytometry data analysis using machine learning'. H Hennig , P Rees , T Blasi , L Kamentsky , J Hung , D Dao , A E Carpenter , A Filby . *Methods* 2017. 112 p. .

[Pesapane et al. ()] 'Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine'. F Pesapane , M Codari , F Sardanelli . *EurRadiol Exp* 2018. 2 (1) p. 35.

[Mulrane et al. (2008)] *Automated image analysis in histopathology: a valuable tool in medical diagnostics. Expert VerMolDiagn*, L Mulrane , E Rexhepaj , S Penney , J J Callanan , W M Gallagher . 2008 Nov. 8 p. .

[Loukas et al. ()] 'Breast cancer characterization based on image classification of tissue sections visualized under low magnification'. C Loukas , S Kostopoulos , A Tanoglidi , D Glotsos , C Sfikas , D Cavouras . *Comput Math Methods Med* 2013. p. 829461.

[Breast Cancer Facts Figures Atlanta ()] 'Breast Cancer Facts & Figures'. *Atlanta* 2019-2020. 2019. American Cancer Society INC.

[Han et al. (2017 23)] 'Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model'. Z Han , B Wei , Y Zheng , Y Yin , K Li , S Li . *Sci Rep* 2017 23. 7 (1) p. 4172.

[Lamprecht et al. ()] 'CellProfiler: free, versatile software for automated biological image analysis'. M R Lamprecht , D M Sabatini , A E Carpenter . *Biotechniques* 2007. 42 (1) p. .

[Carpenter et al. (2006)] 'CellProfiler: image analysis software for identifying and quantifying cell phenotypes'. A E Carpenter , T R Jones , M R Lamprecht , C Clarke , I H Kang , O Friman , D A Guertin , J H Chang , R A Lindquist , J Moffat , P Golland , D M Sabatini . *Genome Biol* 2006. 2006 Oct 31. 7 (10) .

[Araújo et al. ()] 'Classification of breast cancer histology images using Convolutional Neural Networks'. T Araújo , G Aresta , E Castro , J Rouco , P Aguiar , C Eloy , A Polónia , A Campilho . *PLoS One* 2017. 12 (6) p. 177544.

[Chen et al. ()] 'Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: A review'. J M Chen , Y Li , J Xu , L Gong , L W Wang , W L Liu , J Liu . *Tumour Biol* 2017. p. 1010428317694550.

[Misselwitz et al. ()] 'Enhanced CellClassifier: a multi-class classification tool for microscopy images'. B Misselwitz , G Strittmatter , B Periaswamy , M C Schlumberger , S Rout , P Horvath , K Kozak , W D Hardt . *BMC Bioinformatics* 2010. 11 p. 30.

[Incidência do Câncer no Brasil ()] *Incidência do Câncer no Brasil*, 2018. 2017. INCA, Brasil; Rio de Janeiro: INCA.

[Singh et al. (2014)] 'Increasing the Content of High-Content Screening An Overview'. S Singh , A E Carpenter , A Genovesio . *J Biomol Screen* 2014 Jun. 19 (5) p. .

[Lenz et al. ()] D Lenz , L S Gasparini , N D Macedo , E F Pimentel , M Fronza , V L Junior , W S Borges , E R Cole , T U Andrade , Endringerdc . *vitro cell viability by CellProfiler ® software as equivalent to MTT assay*, 2017. 13 p. 365.

[Sommer and Gerlich (2013)] 'Machine learning in cell biology -teaching computers to recognize phenotypes'. C Sommer , D W Gerlich . *J Cell Sci* 2013 Dec 15. 126 p. . (Pt 24)

[Wernickmn et al. (2010)] 'Machine Learning in Medical Imaging'. ; Wernickmn , Y; Yang , J G Brankov , G; Yourganov , Strother , Sc . *IEEE Signal Processing Magazine* 2010 Jul. 27 (4) p. .

[Lu et al. ()] 'Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers'. C Lu , Romo-Buchelid , X Wang , A Janowczyk , S Ganesan , Gilmoreh , D Rimm , A Madabhushi . *Lab Invest* 2018. 98 (11) p. .

[Ching et al.] 'Opportunities and obstacles for deep learning in biology and medicine'. T Ching , D S Himmelstein , B K Beaulieu-Jones , A A Kalinin , B T Do , G P Way , E Ferrero , P M Agapow , M Zietz , M M Hoffman , W Xie , G L Rosen , B J Lengerich , J Israeli , J Lanchantin , S Woloszynek , A E Carpenter , A Shrikumar

, J Xu , E M Cofer , Lavenderca , S C Turaga , A M Alexandari , Z Lu , D J Harris , D Decaprio , Qiy , A Kundaje , Y Peng , L K Wiley , Mhs Segler , S Bocasm , S J Swamidass , A Huang , A Gitter , C S Greene . *J R Soc Interface* 15 (141) p. 20170387.

[Yu et al. (2016)] ‘Predicting non-small cell lung cancer prognosis by fully automated microscopic pathologyimage features’. K H Yu , C Zhang , G J Berry , R B Altman , C Ré , D L Rubin , M Snyder . *Nat Commun* 2016 Aug 16. 7 p. 12474.

[Whitney et al. (2018)] ‘Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer’. J Whitney , G Corredor , A Janowczyk , S Ganesan , Doyles , J Tomaszewski , M Feldman , H Gilmore , Anant Madabhushi . *BMC Cancer* 2018 May 30. 18 (1) p. 610.

[Eulenberg et al. ()] ‘Reconstructing cell cycle and disease progression using deep learning’. P Eulenberg , N Köhler , T Blasi , A Filby , A E Carpenter , Paul Rees , F J Theis , F A Wolf . *Nat Commun* 2017. 8 p. 463.

[Buzin et al. ()] ‘Replacement of specific markers for apoptosis and necrosis by nuclear morphology for affordable cytometry’. R Buzin , F E Pinto , K Nieschke , A Mittag , De Andrade , T U Endringer , D C Tarnok , A Lenz , D . *Journal of Immunological Methods* 2015. 1 p. .

[Jones et al.] *Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning*, T R Jones , A E Carpenter , M R Lamprecht , J Moffat , S J Silver , J K Grenier , A B Castoreno , U S Eggert , D E Root , P Golland , D M Sabatini .

[Xu et al. (2016)] ‘Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images’. J Xu , X Q Lei , H Gilmore , J Wu , J Tang , A Madabhushi . *IEEE Trans Med Imaging* 2016 Jan. 35 (1) p. .

[Beck et al.] ‘Systematic analysis of breast cancer morphology uncovers stromal features associated with survival’. A H Beck , A R Sangoi , S Leung , R J Marinelli , T O Nielsen , M J Van De Vijver , R B West , M Van De Rijn , D Koller . *SciTransl Med* 201 (108) p. .

[Hitchcock ()] ‘The future of telepathology for the developing world’. C L Hitchcock . *Arch Pathol Lab Med* 2011. 135 (2) p. .