Global Journal of Medical Research: C
Microbiology and Pathology
Volume 23 Issue 3 Version 1.0 Year 2023
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 2249-4618 & Print ISSN: 0975-5888

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

By Paulo César Ribeiro Boasquevisque, Robson Dettmann Jarske, Célio Siman Mafra Nunes, Isabela Passos Pereira Quintaes, PhD, Samuel Santana Sodré & Dominik Lenz, PhD

Universidade Vila Velha

Abstract- *Aim:* 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.

Keywords: *breast cancer; image analysis; machine learning; cellular diagnosis; histological malignancy grade.*

GJMR-C Classification: *LCC: RC280.B8*

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

Automated Diagnosis of Breast Cancer

Paulo César Ribeiro Boasquevisque [α], Robson Dettmann Jarske [σ], Célio Siman Mafra Nunes [ρ], Isabela Passos Pereira Quintaes, PhD [Ѡ], Samuel Santana Sodré [¥] & Dominik Lenz, PhD [§]

Abstract- Aim: 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. Regarding the automated classification of the histological degree of malignancy, the Kappa index among the analyzers was 0.55, directly correlating with the frequency of presentation of each graduation group in the analyzed sample.

Conclusion: This study stands out as pioneering research using free access software to diagnose the histological grade in breast cancer and demonstrates that the automated analysis of histopathological parameters is feasible for this purpose.

Keywords: *breast cancer; image analysis; machine learning; cellular diagnosis; histological malignancy grade.*

I. Introduction

Following 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 al., 2017, Xu et al., 2016, Romo-Bucheli et al., 2017, Lu et al., 2018).

Intending to offer agility and safety throughout the diagnosis of diseases, Artificial Intelligence has been increasingly used as a support tool in recent years (Wernick et al., 2010, Mulrane et al., 2008, Jones et al., 2009, Hitchkock et al., 2011, Sommer et al., 2013, Singh et al., 2014., Buzin et al., 2015, Vu et al., 2016, Dordea et al., 2016, Yu et al., 2016, Hennig et al., 2017, Eulenberg et al., 2017, Pesapane et al., 2018, Loukas et al., 2013, Ching et al., 2018).

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).

Author α ρ: Universidade Vila Velha, Espirito Santo, Brazil.
Author σ Ѡ ¥: Universidade Federal Espirito Santo, Brazil.
Corresponding Author §: Universityof Vila Velha. Av. Comissário José Dantas de Melo, n 21. Boa Vista -Vila Velha ES CEP 29102-920. Brazil.
e-mail: dominik.lenz@gmail.com

II. Materials e Methods

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 Research Project received a favorable opinion from the Human Research Ethics Committee of the Cassiano Antônio de Moraes University Hospital of the Federal University of Espírito Santo under No. 2,014,675 of 12/04/2017 and from the Research Ethics Committee on the University of Vila Velha under number 2020,954 of 4/18/2017.

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 best-preserved 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.

c) Loading images to CellProfiler

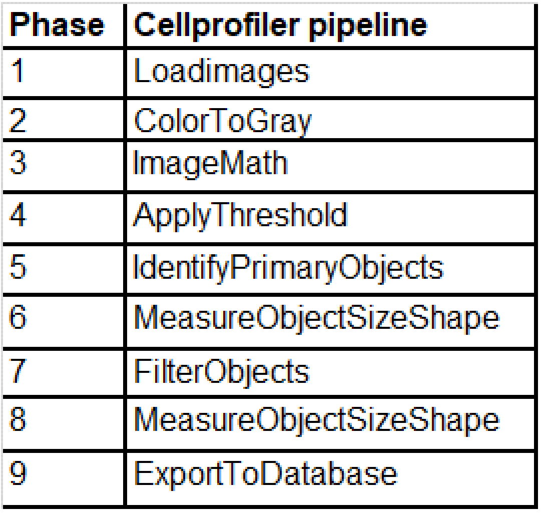
Out of 4,480 digitalized photographs in the 40-fold magnification, after their upload to the CellProfiler program, only the artifact-free images were maintained and recognized as adequate by this image analysis program., Therefore, 1937 images were transferred to the CellProfiler software and submitted to its algorithm, generated for each digitized image with 47 quantitative parameters, called attributes.

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.

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.



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

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.

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).

ii. *Phase 2 – Color to Gray*

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

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 1c).

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.

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 1d).

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.

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 1e).

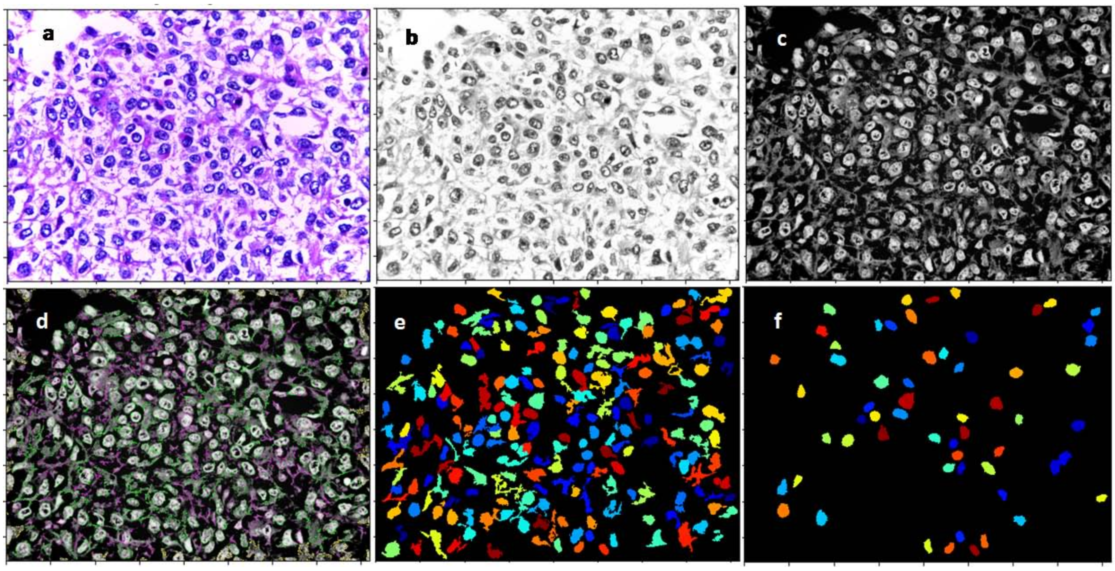


Figure 1a: Original image of breast cancer tissue

Figure 1b: Figure 1a converted to greyscale

Figure 1c: Figure 1b with inverted intensities

Figure 1d & e: Initial identification of nuclei

Figure 1f: Remaining objects after filtering for subsequent analysis

viii. *Phase 8 – Measure Object Size and Shape*

After applying the image filter and eliminating artifactual changes, a new measurement of the primary objects (cell nuclei) attributes was performed.

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.

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.

III. 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. 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 1, while the outcomes comparing the pathological and the automated diagnoses are shown in Table 2. A scatter plot of the automated classification resulted from machine learning is exhibited in Figure 2.

Table 1: Results of the malignancy classification based on image analysis and subsequent classification based on machine learning

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

Table 2: Results of the comparison between pathological and automated analysis

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

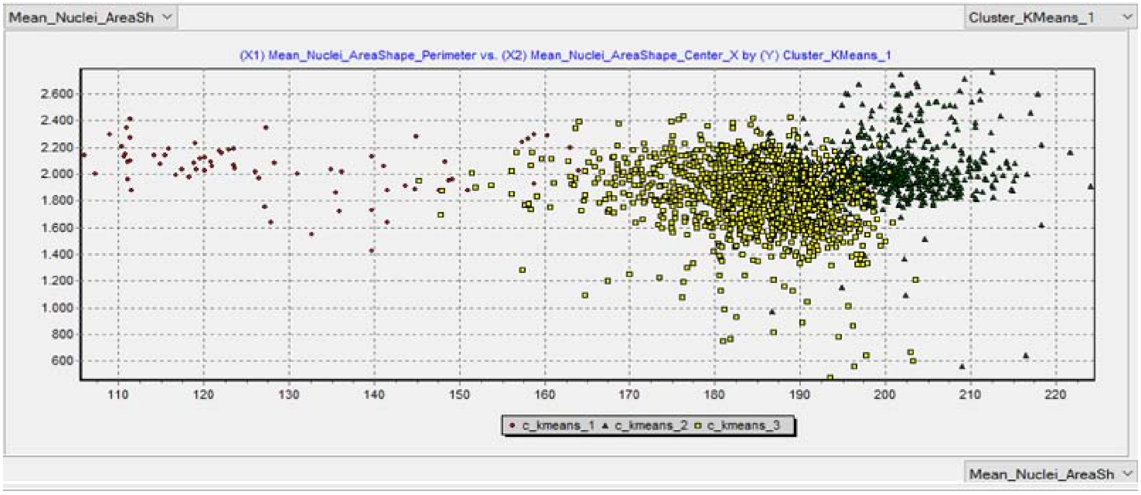


Figure 2a: Classification of malignancy using the tubular score

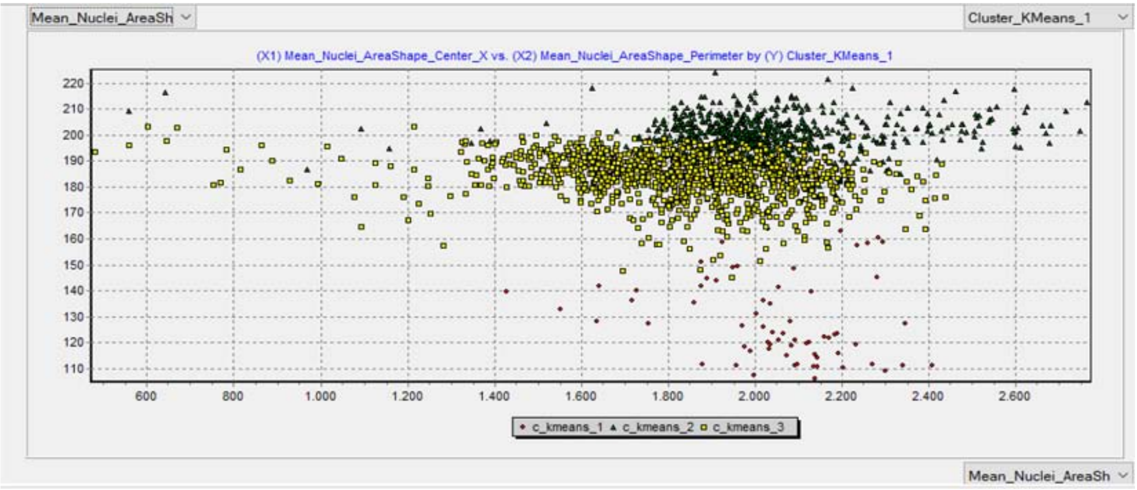


Figure 2b: Classification of malignancy using the nuclear score

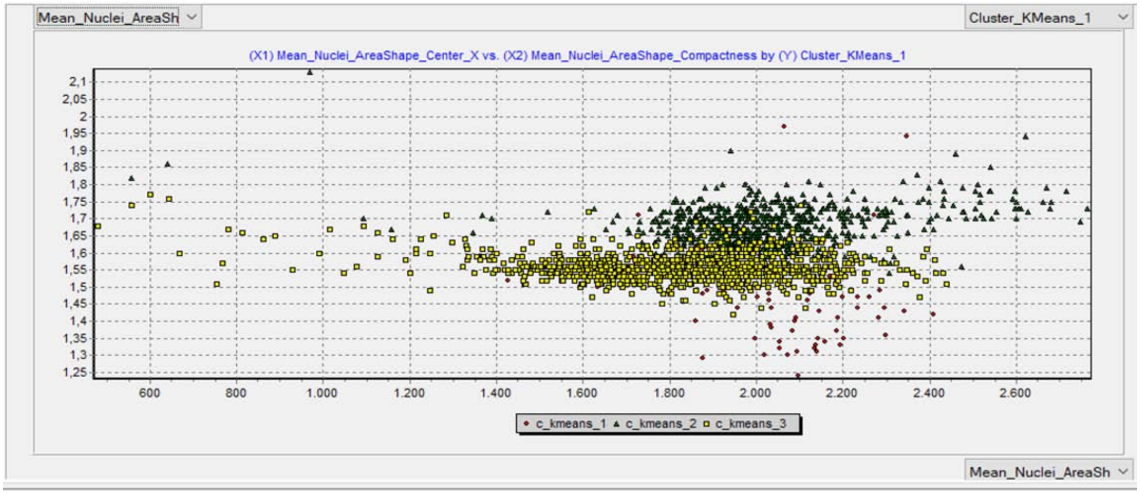


Figure 2c: Classification of malignancy using the mitotic index

Regarding the nuclear score automated classification, the Kappa value represents a sufficient result, while the other analyzed parameters (nuclear score, mitotic index, and histological grade) are considered weak. However, the present study is a pilot study, and further studies are needed to bring more precise results to light.

V. 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.

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.

Bibliographic References

1. American Cancer Society. Breast Cancer Facts & Figures 2019-2020. American Cancer Society INC. Atlanta, 2019.
2. Araújo T, Aresta G, Castro E, Rouco J, Aguiar P, Eloy C, Polónia A, Campilho A. *Classification of breast cancer histology images using Convolutional Neural Networks.* PLoS One. 2017; 12(6): e0177544.
3. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, West RB, van de Rijn M, Koller D. *Systematic analysis of breast cancer morphology uncovers stromal features associated with survival.* SciTransl Med. 201; 3(108): 108ra113.
4. Buzin A R, Pinto F E, Nieschke K, Mittag A, De Andrade T U, Endringer D C, Tarnok A, Lenz D. *Replacement of specific markers for apoptosis and necrosis by nuclear morphology for affordable cytometry.* Journal of Immunological Methods, v. 1, p. 1-6, 2015.
5. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, Guertin DA, Chang JH, Lindquist RA, Moffat J, Golland P, Sabatini DM. *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome Biol. 2006; 7(10): R100. Epub 2006 Oct 31.
6. Chen JM, Li Y, Xu J, Gong L, Wang LW, Liu WL, Liu J. *Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: A review.* Tumour Biol. 2017:1010428317694550.
7. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W, Rosen GL, Lengerich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, LavenderCA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, QiY, Kundaje A, Peng Y, Wiley LK, Segler MHS, BocaSM, S.Swamidass SJ, Huang A, Gitter A, Greene CS. *Opportunities and obstacles for deep learning in biology and medicine.* J R Soc Interface; 15(141): 20170387
8. Dordea AC, Bray MA, Allen K, Logan DJ, Fei F, Malhotra R, Gregory MS, Carpenter AE, Buys ES. *An open-source computational tool to automatically quantify immunolabeled retinal ganglion cells.* Exp Eye Res. 2016; 147:50-56.
9. Eulenberg P, Köhler N, Blasi T, Filby A, Carpenter AE, Paul Rees, Theis FJ, Wolf FA. . *Reconstructing cell cycle and disease progression using deep learning.* Nat Commun.2017;8: 463.
10. Han Z, Wei B, Zheng Y, Yin Y, Li K, Li S. *Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model.* Sci Rep. 2017 23; 7(1): 4172.
11. Hennig H, Rees P, Blasi T, Kamentsky L, Hung J, Dao D, Carpenter AE, Filby A. *An open-source solution for advanced imaging flow cytometry data analysis using machine learning.* Methods. 2017; 112:201-210.
12. Hitchcock CL. The future of telepathology for the developing world. Arch Pathol Lab Med. 2011; 135(2): 211-4.
13. Instituto Nacional de Câncer (INCA, Brasil). Estimativa 2018. Incidência do Câncer no Brasil. Rio de Janeiro: INCA, 2017.
14. Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, Grenier JK, Castoreno AB, Eggert US, Root DE, Golland P, Sabatini DM. *Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning.* ProcNatlAcadSci U S A. 2009 10; 106(6): 1826-31.
15. Lamprecht MR, Sabatini DM, Carpenter AE. *CellProfiler: free, versatile software for automated biological image analysis.* Biotechniques. 2007; 42(1): 71-5.
16. Lenz D, Gasparini LS, Macedo ND, Pimentel EF, Fronza M, Junior VL, Borges WS, Cole ER, Andrade TU, EndringerDC. *In vitro cell viability by CellProfiler ® software as equivalent to MTT assay.* Pharmacognosy Magazine, v. 13, p. 365, 2017.

17. Loukas C, Kostopoulos S, Tanoglidi A, Glotsos D, Sfikas C, Cavouras D. *Breast cancer characterization based on image classification of tissue sections visualized under low magnification.* Comput Math Methods Med. 2013; 829461.
18. Lu C, Romo-BucheliD, Wang X, Janowczyk A, Ganesan S, GilmoreH, Rimm D, Madabhushi A. Nuclear shape and orientation features from H&E images predict survival in early-stage estrogen receptor-positive breast cancers. Lab Invest. 2018; 98(11): 1438–1448.
19. Misselwitz B, Strittmatter G, Periaswamy B, Schlumberger MC, Rout S, Horvath P, Kozak K, Hardt WD. *Enhanced CellClassifier: a multi-class classification tool for microscopy images.* BMC Bioinformatics. 2010; 11:30.
20. Mulrane L, Rexhepaj E, Penney S, Callanan JJ, Gallagher WM. *Automated image analysis in histopathology: a valuable tool in medical diagnostics.* Expert VerMolDiagn. 2008 Nov; 8(6): 707-25.
21. Pesapane F, Codari M, Sardanelli F. *Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine.* EurRadiol Exp. 2018; 2(1): 35.
22. Romo-Bucheli D, Janowczyk A, Gilmore H, Romero E, Madabhushi A. *A Deep Learning Based Strategy for Identifying and Associating Mitotic Activity with Gene Expression Derived Risk Categories in Estrogen Receptor Positive Breast Cancers.* Cytometry A. 2017 Jun; 91(6): 566–573.
23. Singh S, Carpenter, AE, Genovesio A. *Increasing the Content of High-Content Screening An Overview.* J Biomol Screen. 2014 Jun; 19(5): 640–650.
24. Sommer C, Gerlich DW. *Machine learning in cell biology - teaching computers to recognize phenotypes*. J Cell Sci. 2013 Dec 15; 126(Pt 24): 5529-39.
25. WernickMN; Yang Y; Brankov JG; Yourganov G; Strother. SC. *Machine Learning in Medical Imaging.* IEEE Signal Processing Magazine, 2010 Jul; 27(4): 25–38.
26. Whitney J, Corredor G, Janowczyk A, Ganesan S, DoyleS, Tomaszewski J, Feldman M, Gilmore H, Anant Madabhushi. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. BMC Cancer. 2018 May 30; 18(1):610.
27. Xu J, Lei XQ, Gilmore H, Wu J, Tang J, Madabhushi A. *Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images.* IEEE Trans Med Imaging. 2016 Jan; 35(1): 119–130.
28. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, Snyder M. *Predicting non-small cell lung cancer prognosis by fully automated microscopic pathologyimage features.* Nat Commun. 2016 Aug 16; 7: 12474.