

ORIGINAL PAPER

Methodology for structuring a dataset of cervical cell images for the study of malignancy changes associated in conventional Pap Test

Ramon Adrian Salinas Franco ¹, Guilherme P. Coelho ², Paulo S. Martins ²,
Jeannete L. Enciso², Marco A. G. Carvalho ²

¹Federal University of West Bahia – UFOB, ²School of Technology – UNICAMP, ³Foundation University of Health Sciences – FUCS

*radrian20@gmail.com; gpcelho@unicamp.edu.br; paulo@ft.unicamp.edu.br; pajleon@fucsalud.edu.co; magic@ft.unicamp.edu.br;

Received: 2021-05-14. Revised: 2022-04-05. Accepted: 2022-04-13.

Abstract

The Pap test is a screening procedure used worldwide for the diagnosis of cervical cancer. The Malignancy Associated Changes (MAC) are slight morphological and textural changes that take place in the chromatin of the cell nucleus. Their study allows the early detection of several types of cancers. However, the number of data set for MAC studies is limited by the technical aspects that must be followed. In this work, we created a data set for the study of MAC in images of cervical cells from conventional cytology exams. To extract the characteristics of the nuclei, the color channels were separated by the contribution of Hematoxylin and Orange G6 dye, which identified morphological, illumination and texture characteristics. As a result, 1186 nuclei of patients either with some type of anomaly or without anomalies were obtained. To separate the regions of the nuclei, the non-linear regression function and the *Watershed* algorithm with an accuracy of 85.6 % were used. This data set contains adequate data to complement future work in MAC study of cervical cells from conventional pap smears.

Keywords: Image Processing; Watershed; Malignancy Associated Changes; Data set; Papanicolaou Test.

Resumo

As Alterações Associadas à Malignidade (MAC) são pequenas alterações morfológicas e texturais que ocorrem na cromatina do núcleo da célula. Seu estudo permite a detecção precoce de vários tipos de câncer. No entanto, o número de conjuntos de dados para estudos de MAC é limitado pelos aspectos técnicos que devem ser seguidos. Neste trabalho, criamos um conjunto de dados para o estudo do MAC em imagens de células cervicais de exames convencionais de citologia. Para extrair as características dos núcleos, os canais de cores foram separados pela contribuição da Hematoxilina e do Corante Laranja G6, que identificaram características morfológicas, de iluminação e textura. Como resultado, foram obtidos 1186 núcleos de pacientes com algum tipo de anomalia ou sem anomalia. Para separar as regiões dos núcleos, foram utilizados a função de regressão não linear e o algoritmo *Watershed* com precisão de 85,6 %. Este conjunto de dados contém dados adequados para complementar trabalhos futuros no estudo de MAC de células cervicais de esfregaços convencionais de Papanicolaou.

Palavras-Chave: Processamento de imagem; Watershed; Alterações associadas a malignidade; Conjunto de dados; Teste de Papanicolaou.

1 Introduction

The Papanicolaou test (abbreviated as Pap test, also known as Pap smear, cervical smear, or smear test) is a method of cervical screening that involves the removal of human cells, which are spread on a glass slide and colored. The coloration allows the examination under a microscope of any abnormality that indicates a precancerous stage (Paraná et al., 2022). For this, a high number of dyes are used in the procedure. Some of the dyes used are Harris Hematoxylin ($C_{16}H_{14}O_6$), that allows reveal the nuclei of the cells present in the sample; Orange G6 ($C_{16}H_{10}N_2O_6$), which marks DNA molecules and polychrome EA 36, which is used to label neoplastic cells (Castro et al., 2004). The computational studies of digital images depend directly on the adequate coloration of the cytological samples, for the adequate creation of a data set (Tzoannou, 2013).

A data set is a collection of elements contained in a database table, where each column of the table represents a particular variable, and each row represents a particular member of the data set in question. A data set contains the values for each of the variables, such as the area and texture of an object, which correspond to each member of the data set.

A data set is built with a label that contains relevant information and consists of: *Tag* – identification label of the values contained in the tables, and *RV* – Representation value that indicates the type of stored data. Fig. 1 represents how a data set is composed (Chabriaia and Gibaud, 2004).

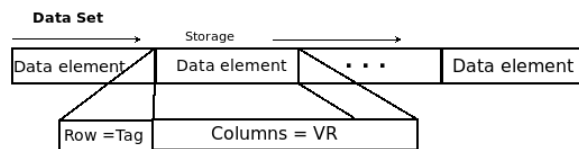


Figure 1: Representation of a conventional data set. Source: Adapted from Chabriaia and Gibaud (2004)

It may be seen that a data set requires two structures. The first is the *Tag*, which contains the information that identifies the data. The other structure contains the value corresponding to the data type *RV*, which can be a string, an integer, a boolean, a float or a binary value (Trujillo et al., 2010).

The cervical cell data set are used for human Papillomavirus studies (Franco et al., 2018). Among these studies we find the identification of MAC for the premature detection of cervical cancer, which is a technique that requires the use of images of cell nuclei with a special treatment. Due to the complexity of extracting this type of images from conventional pap smears, this type of data set is limited. The objective of this work is to create a data set of cervical cells images for the study of MACs using digital image processing techniques.

The data set proposed by this work presents two main contributions. First, it is the first set of data collected for the extraction of nuclei, for MAC studies, from the coloration of the pap smears. Secondly, only the

characteristics of the nuclei were extracted, disregarding the characteristics of the cytoplasm. This decision was made considering that the data extracted from the nuclei are sufficient and adequate to complement MAC studies in conventional pap smears on glass slides. Unlike other works that extracted their data from the cytoplasm and its nuclei.

This work is organized as follows: In Section 2 we address related work. The proposed approach, the fundamentals and details of each step of building the data set are presented in Section 3. Section 4 describes computational procedures to analyze segmented nuclei regions. Finally, our conclusions are presented in Section 5.

2 Related Work

Work carried out at both the hospital and the engineering department of the University of Herlev (Denmark), offered one of the first data set of images of cervical cells from the conventional pap smear (Ampazis et al., 2004). This data set allowed the study and development of computational techniques such as Hard C-means, Wavelet-based, ANFIS neuro fuzzy, Inductive machine learning, Efficient Second Order Neural and Genetic Algorithm based feature (Ampazis et al., 2004, Jantzen and Dounias, 2006, Mitra et al., 1998, Mueen et al., 2008). A feature of this database is that images of conventional human papillomavirus (HPV) cells were used, focusing on the measurement of the morphological aspects of the nucleus and the area of the cytoplasm (Franco et al., 2016). Currently, this data set is discontinued.

A data set made available by the Lancaster University (UK) offers a database of images of cervical cells with a wide diversity and an excellent quality, which were generated from samples of the liquid method (ThinPrep, HOLOGIC Inc., Bedford, USA) (Halliwell et al., 2017). This database is recent and it was developed for researchers who use high-quality images and genetic markers.

Previous image segmentation work on Circular shape constrained Fuzzy Clustering (CiscFC) (Saha et al., 2017), Superpixel-based Markov random field (MRF) (Zhao et al., 2016), Level Set Binário-Hierárquico (Braga et al., 2015) and Nominated Texture (Jayasingh and Stephen, 2014) generated significant advances towards the understanding on how to identify cells that are already affected by some type of anomaly.

The data set of cervical images allow cytotechnicians to identify different types of cell abnormalities as well as the advancement of digital image processing techniques applied to the diagnosis of conventional cervical cancer (Franco, 2020).

Unconventional techniques to diagnose cervical cancer, known as MAC, appeared in the late 70s (Nieburgs, 1967). This method consists of detecting anomalies in the nuclei of the cells with the purpose of making a premature diagnosis with the study of the DNA structure (Finch, 1971). Computational studies developed for decades offered 400 descriptors for the study of the MAC technique (MacAulay et al., 1995). However, the creation of a data set of conventional pap smear samples focused only on the study of the morphological descriptors instead of the

identification of the chromatin, due to the technological limitations of that time. Another limiting factor was that this technique depends on the identification of descriptors directly extracted from the nuclei (to analyze its chromatin), which required specialized (and not easily available) personnel for the collection and digitization of the cervical images (Kemp et al., 1997). Fortunately, these limitations were overcome, thus allowing the contribution of a new data set of cervical cells, such as proposed in our work.

Work by Bengtsson and Malm (2014) proposed a methodology to optimize the cervical cell imaging process for MAC studies. The process starts using a 20X magnification objective lens with an Numerical Aperture (NA) of 0.75 (or higher), a monochromatic camera using a narrow band interference filter with a central passage band of 570 nm. In their methodology, it is recommended to register a stack of 10 foci per image (Bengtsson and Malm, 2014). The idea behind these steps is to efficiently identify the chromatin in the nuclei. The database they used was based on high-quality images acquired by the *ThinPrep* method with pap smears performed in India (Moshavegh et al., 2012). Fig. 2 shows an example of a cervical-cell image acquired from this methodology, where the black spots are the nuclei, the surrounding dark gray area is the cytoplasm and the light gray is the substrate.

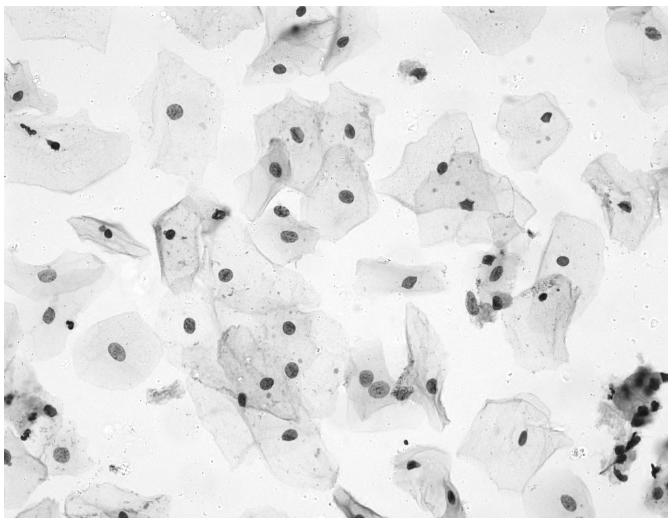


Figure 2: Papanicolaou cell images acquired by using the methodology presented in Bengtsson and Malm (2014)

In South America and specifically in Brazil, the liquid method is not used in the cytological examination. Being among the largest in the world, the Brazilian health system uses a conventional cytology method due to its relative low costs and easy implementation. This reality has not been changing rapidly to a large extent due to the technological and economic borders of Latin countries (da Silva et al., 2022).

Previous work by Franco et al. (2022), in digital images of conventional cytology cells presented a methodology

for segmenting low-quality images using non-linear regression, K-means and the Watershed Transform (Franco, 2019), with favorable results (Fig. 3). In the results of Fig. 3, it may be seen that the nuclei of the cells were clearly identified (despite the complexity presented by the type of images), which presented the opportunity for the creation of a data set for the study of MAC.

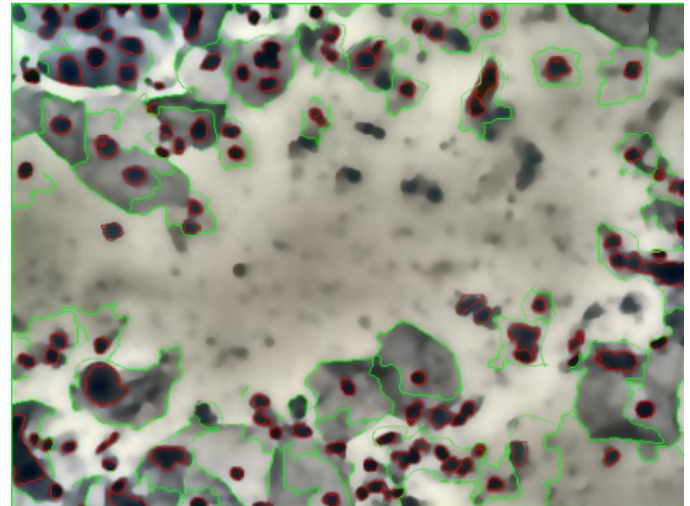


Figure 3: Segmentation of the conventional human papilloma exam. Source: (Franco, 2019)

In this work, a data set is presented with a *Tag* structure that identifies the data and a *RV* structure that identifies the type of data. The structure of specific data was divided into two types: demographic data and data extracted from morphology, texture and illumination of the images.

3 Methods

In this section, we describe the activities carried out and the equipment used for the recruitment of patients, the acquisition of conventional pap smears and the digitalization of the images. In addition, we also address the steps for coloring, preprocessing, segmentation and annotation of the digitized cells for the creation of the data set. Fig. 4 illustrates the methodology.

We grouped the proposed approach in three main stages: (i) the collection of cervical cells, where the recruitment of patients and the technical tasks of collection and coloration were carried out; (ii) the digitization of the slides; and (iii) the image processing. In stage three, we performed the computational analysis, the data record and the creation of the data set files. The next subsection describes each one of the steps presented in Fig. 4.

This study was approved by the Research Ethics Committee, a human-research ethics committee (CAAE approval number: 71277217.6.0000.5404) (Ministério da Saúde, 2018), and thus it was conducted in accordance with the principles of applicable national and local regulations. All patients gave their written consent before

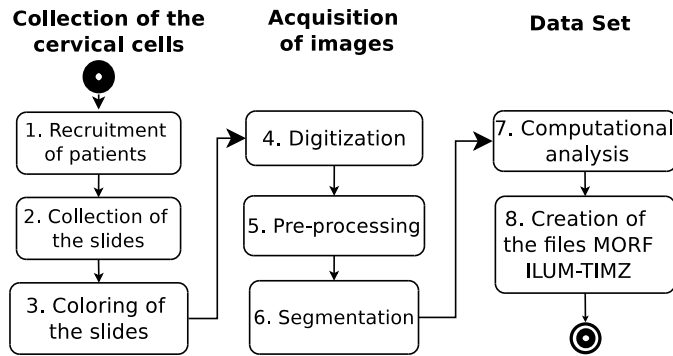


Figure 4: Methodology used in the study

carrying out any procedure specific to the protocol.

3.1 Patient recruitment

A cervical cancer prevention campaign was carried out at the University of Campinas (Brazil), where 71 women were enrolled. For this study, six women were chosen: non-pre-menopausal, non-pregnant and between the ages of 25 and 40 years. Samples diagnosed with normal squamous lesions, atypia in squamous cells of undetermined significance (AC-US), low-grade squamous intraepithelial lesion LEI (changes associated with HPV infection or light dislocation (NIC 1)), and immature squamous metaplasia, were chosen. All samples were collected before treatment. The patients were anonymized and assigned a unique identifier. The criterion to limit the number of slides used in this study is related to the large amount of data to be processed and the computational cost necessary for this process.

3.2 Collection of cervical cells and preparation of the slides

Samples were collected following the protocol established by the World Health Organization (WHO) for the acquisition of conventional pap smears ([Organización Panamericana de la Salud, 2020](#)). The collection was supervised by a specialist from the University Foundation of Health Sciences (FUCS), who has over 30 years of experience in this field.

The characteristics of the patients were collected and recorded in an information system. The data included ethnicity, sexually transmitted infection (STI), background, contraceptive method, procedures in the cervix, if smoker, gestations, menstrual cycle, if infant, birth and slide identification. The medical and gynecological history of each patient was also collected, including the time elapsed since the last sexual intercourse. For each patient, the medical staff collected data on cytology, excluding both women under 18 and those pregnant or menopausal.

For the preparation of slides, each sample was fixed with cytological spray and successively immersed in concentrations of alcohol of 90, 80, 70, and 50% (in water), one minute in each liquid. Each sample was

then stained with Harris Hematoxylin, Hydrochloric Acid and Orange. Finally, each of the plates was mounted for observation under a microscope. The major modification of the procedure was to submerge the samples in the Orange G liquid, adding one extra minute to the protocol specification value in order to generate a larger contrast of the nuclei for subsequent digitization.

3.3 Obtaining the conventional Papanicolaou slides

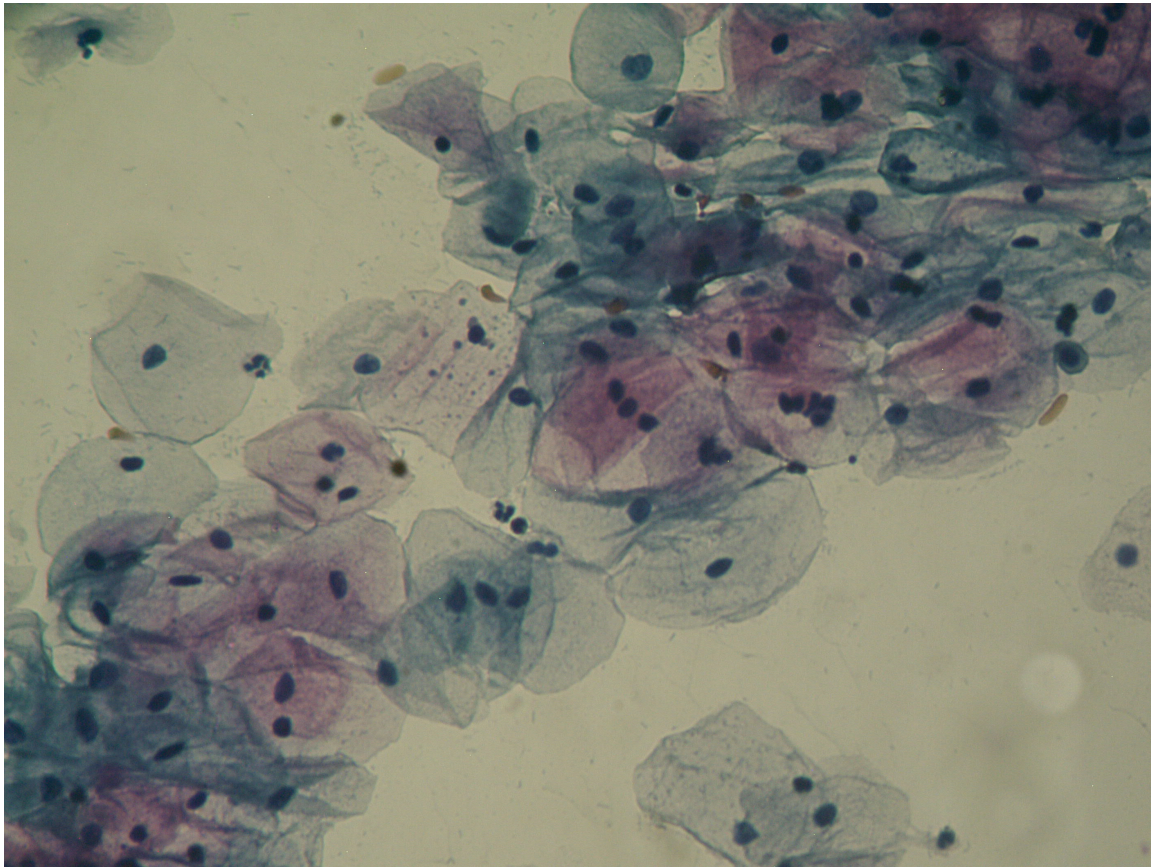
A low-resolution Polaric camera was mounted on a JENAMETD 2 microscope. The images were captured with a 20X optic and with a Numerical Aperture (NA) of 0.4mm. A set of 1000 images was captured and, from that, ten images were randomly selected per slide, for the extraction of the cells from the six selected patients. The cells that were too small or that were insulated were not processed. The experiments were carried out with groups of cell images as performed by pathology specialists ([Fig. 5\(a\)](#)) and not in an insulated way as seen in other publications ([Ampazis et al., 2004](#), [Marinakos et al., 2009](#)).

3.4 Preprocessing

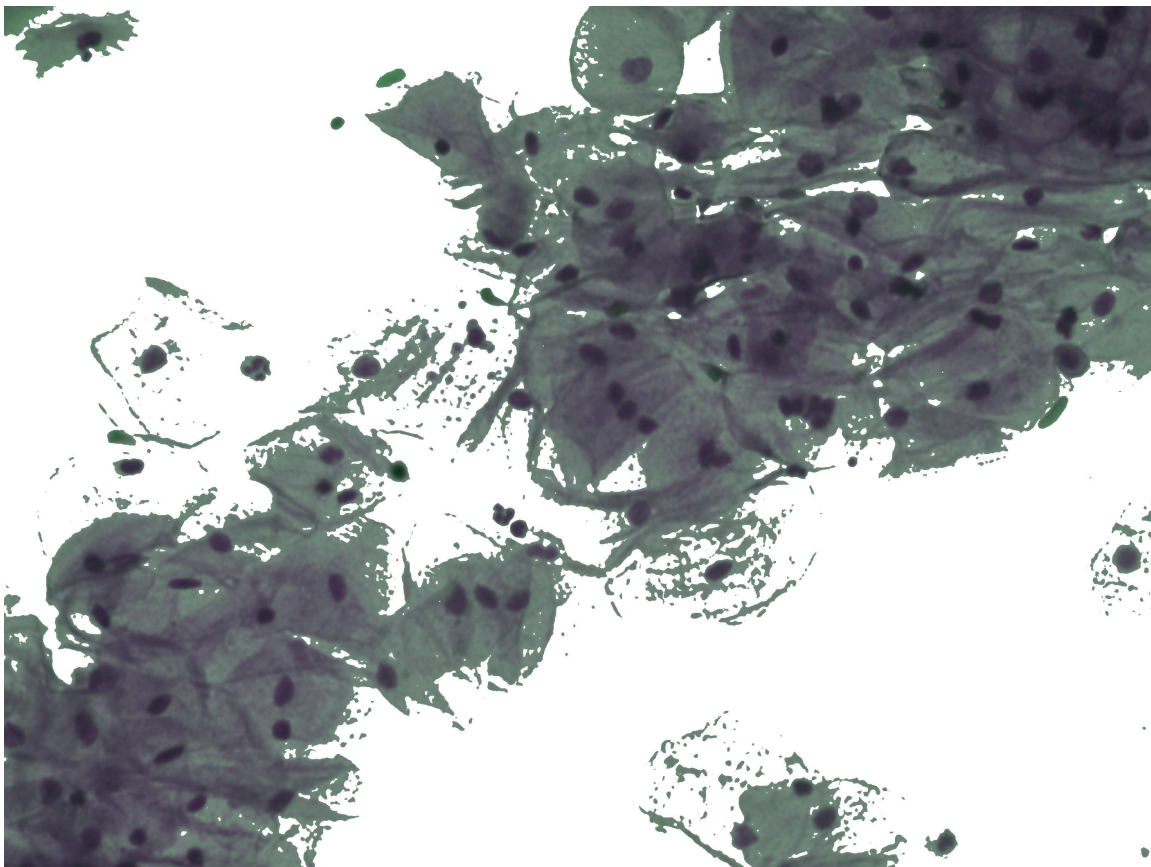
The image presented in [Fig. 5 \(a\)](#) was preprocessed since they presented high content of debris such as mucous, blood and fluids. The methodology proposed by Franco et al. for low-quality images was adopted ([Franco et al., 2016](#)). The correction of the median of the height line on the horizontal and vertical axis was applied. Following the elimination of the general noise of the image, a bilateral filter ([Paris et al., 2008](#)) was employed to eliminate high-frequency noise and maintain the edges of the images. Next, the background of the image was detected using non-linear regression. With this procedure, it was possible to identify the biological materials to be segmented. In [Fig. 5 \(b\)](#), it is possible to see the results.

3.5 Segmentation

After preprocessing of images, the segmentation step was implemented. The images were grouped in seven gray distributions using the *K*-means algorithm. The number of groups was defined according to the segmentation methodology of low-quality images previously mentioned. This methodology allows the determination of the cytoplasm and the nuclei. However, for this study, only the nuclei of the cells were considered. The reason for using *K*-means is to find groups of pixels in the image that could be used as seed by the Watershed Transform, and, therefore, perform the nuclei segmentation. The coordinates of the images captured by the grouping process were dilated to determine some pixel seeds. These pixels were used by the Watershed algorithm to determine the kernels. Each image was dilated by three pixels in all directions to determine the images with the most significant nuclei. The segmented nuclei were cropped and stored in a folder for later computational analysis. [Fig. 6](#) presents some of the segmented nuclei.



(a) Original image



(b) Pre-processed image

Figure 5: Comparison of the original and pre-processed image, where part of the biological material was filtered

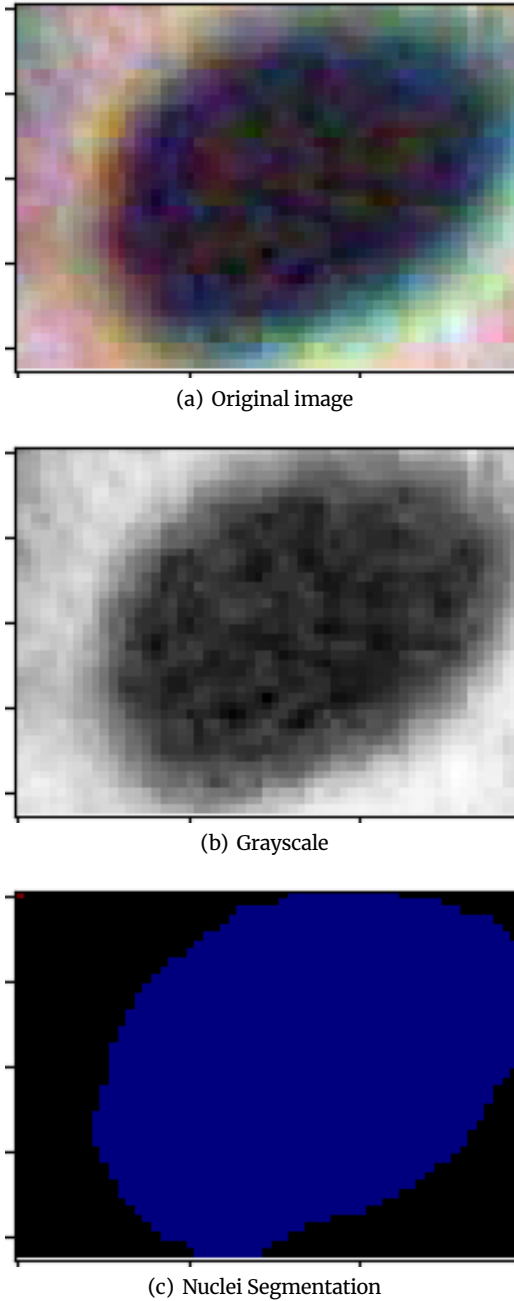


Figure 6: Nuclei's region segmentation process.

4 Computational Analysis

The raw segmented regions were processed by the software *CellProfiler Stable 3.0.0 cell image analysis* for each of the segmented nuclei (Kamentsky et al., 2011). For each cell under analysis, a set of three files was generated, corresponding to three general types of characteristics of the images, i.e. morphological, texture and lighting characteristics.

The regions of the nuclei were separated by dye staining, since the dyes absorb an amount of light in the red, green

and blue channels that proportionally increases specific characteristics in each channel (Ruifrok, 2001). The absorption factor A is described by Lambert Beer's law (Swinehart, 1962), given in Eq. (1).

$$A = \log_{10} \left(\frac{I_0}{I} \right) = \epsilon \ell c \quad (1)$$

where I_0 is the intensity of the outgoing light after passing through the specimen, I is the intensity of the incoming light before passing the specimen, ℓ is the length traversed by the light in the middle, the constant ϵ is the molar absorptivity and c is the concentration of the absorbent in the medium that characterizes each of the RGB channels. Since the gray level values are not linear with the intensity values and cannot be calculated, it is required to define the optical density OD for each RGB channel by means of Eq. (2):

$$OD = \frac{A}{L} \quad (2)$$

where OD is the optical density of the specimen, L is the thickness of the sample and A is the absorbance.

Eq. (1) indicates the linear relationship between the OD of each channel and the concentration of absorbent material. In this case, we use OD to separate the contribution of each color of dye in the image, where each dye has an OD of the chromatin in each of the RGB channels. Fig. 7 presents the result of the Harris Hematoxylin channels and Orange G6 dye separation.

This separation by color contribution allows the subtle detection of chromatin stains that allows the study of MACs in tests performed by conventional cytology, offering valuable information in the construction of data set. After the separation by color contribution, the images were converted to gray scale, for the measurements of the morphological characteristics and illumination of the nuclei. The data record obtained from each of the images is described in the sequence.

4.1 Data records header file

The data record was composed of metadata that inform the demographic data of the images. Table 1 details the characteristics of the six selected patients and this information is part of a *Tag* associated with the demographic data of the images.

4.2 Data set of conventional Pap Smear images

Raw data set are available as CSV files. Each data was numbered to include the identification of the metadata, along with the tag of the images following a sequence. For example, the data file *01-CAP091868-MORE.csv* has the prefix 01 that refers to the sequential numbering of the files with the data from the digitized slide.

Table 2 represents how the data set files were created. For example, CAP091868 refers to the unique identifier of a patient. The four letters that follow this identifier indicate the type of data extracted. MORF is the morphological

Table 1: Demographic data of the slides

	Negative Malignancy	Negative Malignancy	Low-grade Lesion HPV	Low-grade Lesion HPV	Negative Malignancy ASCUS	Negative Malignancy ASGU
Patient identifier	MBP072994	GSD083193	CAP091868	RC072682	NP161194	LC050283
Nucleus	99	62	341	282	110	292
Age	25	25	46	35	25	34
Ethnicity	White	Black	White	Pardo	White	White
STI (Sexually Transmitted Infection)	Not	Not	Not	Not	Not	Not
Contraceptive methods	Pill	Condon	Not	Pill	Pill	Pill
Abortions	0	0	0	0	0	0
Menstrual phase	21	28	28	31	21	28
Squamous Atipocal Cell	Not	Yes	Not	Yes	Yes	Not
Cervical procedures	0	0	0	0	0	1
Gestations	0	0	3	1	0	0
Breast-feed	Not	Not	Not	Not	Not	Not
Births	0	0	3	1	0	0
Appearance Uterus	Healthy	Healthy	Healthy	Healthy	Healthy	Healthy
Previous cytology	Yes	Not	Yes	Yes	Yes	Not

Table 2: Example of description of data set files

Patient identifier	Data type	Description
CAP091868	MORF	Morphological information of nuclei
CAP091868	TEXT	Texture information of nuclei
CAP091868	INTE	Intensity information of nuclei
RC072682	MORF	Morphological information of nuclei

information of the image of an unprocessed nucleus, whereas TEXT refers to the texture values detected in the segmented nucleus. This identification structure refers to the *Tag* data set. INTE covers all the information concerning the intensity of the image.

The analysis of data such as optical density, area, mean diameter, perimeter, and radial distribution was only performed using MORF data. The INTE and TEXT files were included in the data set as a complement and would generally be used for future work.

The values of each of the nuclei have a specific value calculated individually in pixels. The data were manually verified by cytopathology specialists to ensure that the data were correct. Each file extension contains the date the data was collected. It is possible to easily access and work with this data through data mining systems, for example, for its preparation, classification, regression, grouping, association rule mining, or visualization.

5 Remarks and Conclusion

Cervical cancer is a widespread infectious disease that affects thousands of women worldwide every year. Due to the alarming rate, it is of utmost importance that the medical staff is provided with the needed information to further advance research in this field. In particular, medical imaging is relevant since it is used to determine whether or not cancer has spread. Within this context, this work attempted to provide a data set of images consisting of six patients with different characteristics in their diagnoses.

The images presented a large amount of biological detritus and required the pre-processing of images to eliminate unnecessary biological material. A total of 1306

candidate nuclei were identified from the segmentation of the images using the automated methodology. In order to reduce noise in the experiments, a manual and visual filtering was carried out to select the well-defined nuclei (i.e. those that were not overlapping), thus resulting in 1186 images (85.6% of the nuclei).

The methodology to extract the characteristics of the images of the nuclei for the study of MAC proposed by Bengtsson et al. recommends the use of an optic with a numerical aperture of 0.75 mm. In our experiments, an optic of 0.4 mm was used, which slightly limited the definition of the chromatin of the nuclei. However, during the collection of samples, the submersion of the plates in the Hematoxylin dye for an extra minute resulted in images with a more detailed nucleus. The computational separation by the color contribution of Hematoxylin and Orange 6 offered a subtle perspective of chromatin in the images, thus facilitating the study of MAC's. Different variations were observed in the extracted data, in the nuclei of the sheets that presented some type of alteration, specifically in the nuclei area and perimeter. This behavior was expected since nuclear enlargement is a known condition in altered cells.

It is concluded that the methodology of preprocessing and processing of an image offered satisfactory results since it was able to detect most of the nuclei for the data set. This indicates that the use of clustering algorithms such as the *K*-means in combination with the Watershed transform can identify nucleus in low-quality images and difficult segmentation.

In conclusion, from the six patients analyzed, four of the diagnoses were negative for malignancy but detected the presence of ASCUS, which may be an indicator of future lesions. Furthermore, two of the diagnoses were

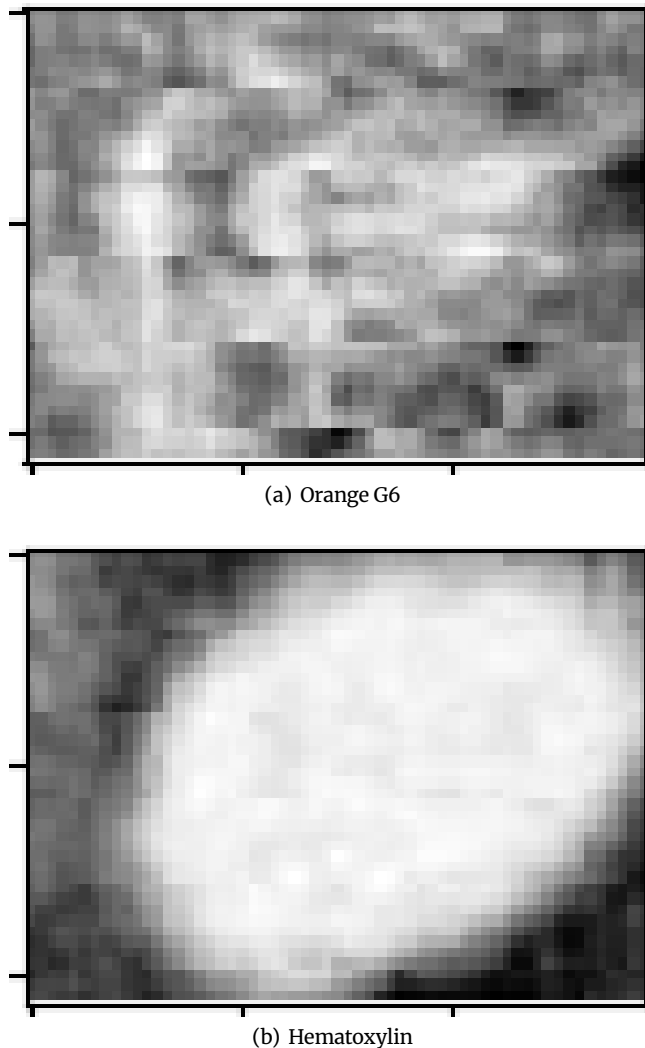


Figure 7: Harris Hematoxylin channels and Orange G6 dye separation: (a) nucleus representation of Orange G6; (b) nucleus in the Hematoxylin coloration.

positive for malignancy as they showed low NIC1 grade lesions. All the results were corroborated by a specialist in cytopathology from the University Health Sciences Foundation, thus validating our procedures and results.

In addition to the diagnosis, this work also contributed with an image database of automatically segmented cervical nuclei along with a set of morphological, textural and illumination information. It may be regarded as a reference for image segmentation through which classification algorithms may be objectively evaluated for the identification of MACs. Over time, we intend that the database grows to cover 10,000 images that include other types of features. This information will be made available to the community so that we may continue with the MACs studies using low-quality images in conventional Pap smear.

Acknowledgments

We thank all the participants in this study. Especially to the researcher Msc. Juan Galindo, the CAISM Women's Hospital, the CECOM Community Center, the FUCS University Health Sciences Foundation, the School of Technology of the University of Campinas - UNICAMP, the Universidade Paulista - UNICEP. This work was carried out during a scholarship supported by the CAPES International Cooperation Program at UNICAMP.

References

- Ampazis, N., Dounias, G. and Jantzen, J. (2004). Pap-Smear Classification Using Efficient Second Order Neural Network Training Algorithms, *Lecture Notes in Artificial Intelligence* **3025**: 230–245. https://doi.org/10.1007/978-3-540-24674-9_25.
- Bengtsson, E. and Malm, P. (2014). Screening for Cervical Cancer Using Automated Analysis of PAP-Smears, *Computational and Mathematical Methods in Medicine* **2014**: 1–12. <http://dx.doi.org/10.1155/2014/842037>. URL: <http://www.hindawi.com/journals/cmml/2014/842037/>
- Braga, A. M., Marques, R. C. P. and Ushizima, D. M. (2015). Segmentação de Células Cervicais Utilizando Level Set Binário-Hierárquico, *Conference on Graphics, Patterns and Images*, 28 (SIBGRAPI), Sociedade Brasileira de Computação, Porto Alegre. Available at http://sibgrapi.sid.inpe.br/col/sid.inpe.br/sibgrapi/2015/07.13.18.25/doc/segmentacao-de-celulas_camera_ready.pdf.
- Castro, M. I., Abratte, O., Barocchi, M. and Musacchio, M. L. (2004). Coloración de Papanicolaou y su importancia en el diagnóstico de las infecciones cervicovaginales, *Acta bioquímica clínica latinoamericana* **38**(2): 199–202. Available at http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S0325-29572004000200008.
- Chabriaes, J. and Gibaud, B. (2004). DICOM, le standard pour le imagerie medicale, *EMC-Radiologie* **1**(6): 577–603. <https://doi.org/10.1016/j.emcrad.2004.09.002>.
- da Silva, D., Garnelo, L. and Herkrath, F. (2022). Barriers to access the pap smear test for cervical cancer screening in rural riverside populations covered by a fluvial primary healthcare team in the amazon, *Int. J. Environ. Res. Public Health* **19**(7): 4193. <https://doi.org/10.3390/ijerph19074193>.
- Finch, R. R. (1971). A classification of nuclear aberration in relation to malignancy associated changes (MAC)., *Acta cytologica* **15**(6): 553–8. Available at <http://www.ncbi.nlm.nih.gov/pubmed/4332996>.
- Franco, R. A. S. (2019). *Optimización de descriptores usados en los cambios asociados a la malignidad en imágenes digitales de células cervicales*, Thesis, University of Campinas, Faculty of Technology.
- Franco, R. A. S. (2020). Database focused on conventional cervical cells (FAPESP – Phase 1). <https://doi.org/10.5281/zenodo.4148444>.

- Franco, R. A. S., Carvalho, M. A. G., Coelho, G. P., Martins, P. and Enciso, J. L. (2018). Dataset of Cervical Cell Images for the Study of Changes Associated with Malignancy in Conventional Pap Test. <https://doi.org/10.5281/zenodo.1409790>.
- Franco, R. A. S., Coelho, G. P., Pedro, P. S. M. and de Carvalho, M. A. G. (2022). Descriptores mac en la detección temprana del cáncer cervical usando técnicas de procesamiento de imágenes, *Revista Cubana de Ciencias Informáticas* 16(1). Available at http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2227-18992022000100001.
- Franco, R. A. S., Martins, P. S. and Carvalho, M. A. G. D. (2016). Cytological Low-Quality Image Segmentation Using Nonlinear Regression, K-means and Watershed, *Computer Science Research Notes - CSRN* 2601 pp. 91–98. Available at <https://dspace5.zcu.cz/bitstream/11025/29535/1/Franco.pdf>.
- Halliwell, D. E., Morais, C. L. M., Siggel-king, M. R. F., Craig, T., Ingham, J., Martin, D. S., Heys, K., Kyrgiou, M., Mitra, A., Paraskevaidis, E., Theophilou, G., Martin-hirsch, P. L., Cricenti, A., Luce, M., Weightman, P. and Martin, F. L. (2017). An imaging dataset of cervical cells using scanning near-field optical microscopy coupled to an infrared free electron laser, *Sci Data* 4(170084): 1–11. <https://doi.org/10.1038/sdata.2017.84>.
- Jantzen, J. and Dounias, G. (2006). Analysis of Pap-Smear Image Data, *Proceedings of the Nature-Inspired Smart Information Systems 2nd Annual Symposium* 305: 10. Available at <https://backend.orbit.dtu.dk/ws/portalfiles/portal/2528109/oersted-dtu2886.pdf>.
- Jayasingh, E. and Stephen, A. (2014). Nominated Texture Based Cervical Cancer Classification, *Computational and Mathematical Methods in Medicine* 2015: 10. <https://doi.org/10.1155/2015/586928>.
- Kamentsky, L., Jones, T. R., Fraser, A., Bray, M.-A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W. and Carpenter, A. E. (2011). Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software., *Bioinformatics (Oxford, England)* 27(8): 1179–80. <http://dx.doi.org/10.1093/bioinformatics/btr095>.
- Kemp, R. A., MacAulay, C., Garner, D. and Palcic, B. (1997). Detection of malignancy associated changes in cervical cell nuclei using feed-forward neural networks., *Analytical Cellular Pathology* 14(1): 31–40. <https://doi.org/10.1155/1997/839686>.
- MacAulay, C., Lam, S., Payne, P. W., LeRiche, J. C. and Palcic, B. (1995). Malignancy-associated changes in bronchial epithelial cells in biopsy specimens., *Analytical and quantitative cytology and histology / the International Academy of Cytology [and] American Society of Cytology* 17(1): 55–61. Available at <http://www.ncbi.nlm.nih.gov/pubmed/7766269>.
- Marinakakis, Y., Dounias, G. and Jantzen, J. (2009). Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification, *Computers in Biology and Medicine* 39(1): 69–78. <https://doi.org/10.1016/j.combiomed.2008.11.006>.
- Ministério da Saúde (2018). Seems Ethics Committee. Available at <http://plataformabrasil.saude.gov.br/login.jsf>.
- Mitra, S., Yang, S. and Kustov, V. (1998). Wavelet-based vector quantization for high-fidelity compression and fast transmission of medical images, *Journal of digital imaging* 11(2): 24–30. <https://doi.org/10.1007/BF03168174>.
- Moshavegh, R., Ehteshami Bejnordi, B., Mehnert, A., Sujathan, K., Malm, P. and Bengtsson, E. (2012). Automated segmentation of free-lying cell nuclei in Pap smears for malignancy-associated change analysis, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2(9): 5372–5375. <http://dx.doi.org/10.1109/EMBC.2012.6347208>.
- Mueen, A., Zainuddin, R. and Baba, M. S. (2008). Automatic multilevel medical image annotation and retrieval, *Journal of digital imaging* 21(3): 290–295. <https://doi.org/10.1007/s10278-007-9070-3>.
- Nieburgs, H. E. (1967). Recent progress in the interpretation of malignancy associated changes (MAC)., *Acta cytologica* 12(6): 445–453.
- Organización Panamericana de la Salud (2020). *Experiencias con la implementación de programas de tamizaje de cáncer cervicouterino basados en la prueba de VPH. Informe de reunión*, Vol. 20, OPS.
- Paraná, V. C., Souza Santos, D., Barreto de Souza Silva, D. I., Lima, G. C., Gois, L. L. and Santos, L. A. (2022). Anal and cervical human papillomavirus genotypes in women co-infected with human immunodeficiency virus: A systematic review, *International journal of STD & AIDS* p. 09564624221076293. <https://doi.org/10.1177/09564624221076293>.
- Paris, S., Kornprobst, P., Tumblin, J. and Durand, F. (2008). Bilateral Filtering: Theory and Applications, *Foundations and Trends® in Computer Graphics and Vision* 4(1): 1–75. <http://dx.doi.org/10.1561/06000000020>.
- Ruifrok, A. C. (2001). Quantification of histochemical staining by color deconvolution, *Anal Quant Cytol Histol* 23: 291–299.
- Saha, R., Bajger, M. and Lee, G. (2017). Circular shape constrained fuzzy clustering (ciscfc) for nucleus segmentation in pap smear images, *Computers in Biology and Medicine* 85: 13 – 23. <https://doi.org/10.1016/j.combiomed.2017.04.008>.
- Swinehart, D. F. (1962). The Beer-Lambert Law, *Journal of Chemical Education* 39(7): 333. <http://dx.doi.org/10.1021/ed039p333>.
- Trujillo, J. P., Rivera, J. H. and Serna, W. (2010). Descripción Del Estándar Dicom Para Un Acceso Confiable a La

Información De Las Imágenes Médicas ., *Scientia et Technica* 45(0122-1701): 289–294. <https://doi.org/10.22517/23447214.347>.

Tzoannou, N. (2013). *Automatic Detection of Blood Vessels in Medical Images*, Thesis, Smt. Kasturbai Walchand College.

Zhao, L., Li, K., Wang, M., Yin, J., Zhu, E., Wu, C., Wang, S. and Zhu, C. (2016). Automatic cytoplasm and nuclei segmentation for color cervical smear image using an efficient gap-search mrf, *Computers in Biology and Medicine* 71: 46 – 56. <https://doi.org/10.1016/j.compbimed.2016.01.025>.