Title

Cric searchable image database as a public platform for conventional pap smear cytology data

Permalink

https://escholarship.org/uc/item/1717m4s3

Journal

Scientific Data, 8(1)

ISSN

2052-4463

Authors

Rezende, Mariana T

Silva, Raniere

Bernardo, Fagner de O

et al.

Publication Date

2021

DOI

10.1038/s41597-021-00933-8

Peer reviewed

scientific data

OPEN

ARTICLE

Cric searchable image database as a public platform for conventional pap smear cytology data

Mariana T. Rezende[1,2,8] ✉, Raniere Silva[3,8] ✉, Fagner de O. Bernardo[3], Alessandra H. G. Tobias[2], Paulo H. C. Oliveira[3], Tales M. Machado[3], Caio S. Costa[3], Fatima N. S. Medeiros[4], Daniela M. Ushizima[5,6,7], Claudia M. Carneiro[1,2] & Andrea G. C. Bianchi[3]

Amidst the current health crisis and social distancing, telemedicine has become an important part of mainstream of healthcare, and building and deploying computational tools to support screening more efficiently is an increasing medical priority. The early identification of cervical cancer precursor lesions by Pap smear test can identify candidates for subsequent treatment. However, one of the main challenges is the accuracy of the conventional method, often subject to high rates of false negative. While machine learning has been highlighted to reduce the limitations of the test, the absence of high-quality curated datasets has prevented strategies development to improve cervical cancer screening. The Center for Recognition and Inspection of Cells (CRIC) platform enables the creation of CRIC Cervix collection, currently with 400 images (1,376 × 1,020 pixels) curated from conventional Pap smears, with manual classification of 11,534 cells. This collection has the potential to advance current efforts in training and testing machine learning algorithms for the automation of tasks as part of the cytopathological analysis in the routine work of laboratories.

Introduction

Cervical cancer is one of the most frequently diagnosed neoplasms and one of the main causes of death from cancer in the female population, and constitutes a significant public health problem worldwide[1]. According to the most recent estimate by the World Health Organization (WHO), it is the fourth most incident cancer among women worldwide, with approximately 342,000 deaths in 2020, and is the leading cause of cancer death in 42 countries[2].

Cervical cancer has one of the best prognosis for prevention and cure, reaching almost 100% of cure when diagnosed early with screening methods[3]. The cytopathological procedure known as Pap smear[4], which is the most widely used test, is relatively inexpensive to perform, and effective in detecting precursor lesions[4].

The discovery that cervical infection by high oncogenic risk human papillomavirus (HPV) genotypes can progress to cervical cancer has led to the advancement of HPV molecular detection tests to screen for this neoplasm[5]. However, the Pap test is still essential, since positivity for oncogenic HPV still requires cytological information[6,7] as a follow-up.

Even with molecular diagnosis, Pap smear is still the screening method used in low-income countries: its replacement by more expensive methods is a distant reality[3]. Even in high-income countries, well-structured implementation of programs using the Pap smear has reduced cervical cancer incidence and mortality rates by up to 65% in the last 40 years[8].

[1]Graduate Program in Biotechnology, Biological Sciences Research Center (NUPEB), Federal University of Ouro Preto, Ouro Preto, MG, Brazil. [2]Cytology Laboratory, Clinical Analysis Department, Federal University of Ouro Preto, Ouro Preto, MG, Brazil. [3]Computing Department, Federal University of Ouro Preto, Ouro Preto, MG, Brazil. [4]Teleinformatics Engineering Department, Federal University of Ceará, Fortaleza, CE, Brazil. [5]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [6]Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA. [7]Berkeley Institute for Data Science, University of California Berkeley, Berkeley, CA, USA. [8]These authors contributed equally: Mariana T. Rezende, Raniere Silva. ✉e-mail: mariana.trevisan@aluno.ufop.edu.br; raniere.gcs@aluno.ufop.edu.br

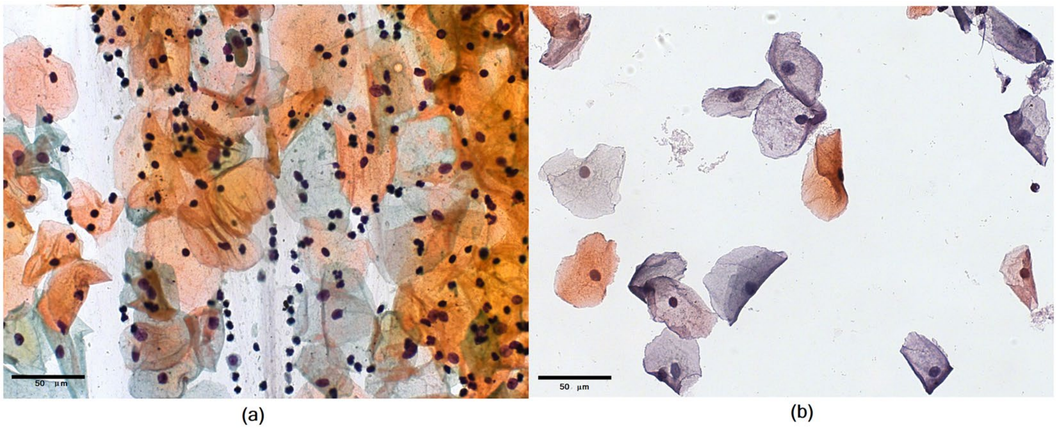

Fig. 1 Illustration of (**a**) conventional cytology and (**b**) liquid-based cytology.

Despite the global use of Pap smear to detect cervical cancer, it has inherent limitations[9], including the fact that it is labor intensive, underscoring the need for strategies that generate more accurate results and with lower rates of false-negative, false-positive and unsatisfactory results. Since visual interpretation is time-consuming, subjective and requires highly specialized human interaction[10–15], increasing efforts over the past few decades have been aimed at developing automated analysis of Pap smear data.

In recent years, machine learning algorithms have been proposed to tackle these challenges. One of the most crucial requirements for an automated system powered by computer vision and machine learning techniques is a collection of hundreds (if not thousands) of high-quality, well-curated digitized images of Pap smears and related metadata. The existing FAIR (Findable, Accessible, Interoperable, and Reusable)[16] data about Pap smears is limited, and without representation of cells with pre-neoplastic alteration of all classes, mostly composed of cut-out cells, synthetic, non-standardized images, and from liquid-based cytology[17–23].

The most widely used FAIR data in Pap smear image classification studies for cervical lesion detection is the Herlev base[5,17,21,22,24]. This database comprises a total of 917 images: each image contains a single cervical cell and is assigned to one of seven classes of a pre-neoplasic lesion. Other FAIR data used in Pap smear image classification studies include the SIPaKMeD[25] database, with 4,049 cells (in 966 images) assigned to one of five categories of cell types, with no pre-neoplastic alteration.

Another issue with most of the databases, including Herlev and SIPaKMeD, is that they are not classified according to the Bethesda System nomenclature, created in 1988 to standardize the terminology of cervicovaginal cytology in order to reduce high variability in the communication of results. The Bethesda System is considered the most used, uniform and reproducible terminology among different pathologists and laboratories[26,27].

A further complication is that many previous works have relied on image collections with cut-out cells and/or unrealistically "clean" images, similar to those in liquid-based cytology[10,20,25,28]. Figure 1 illustrates the drastic difference between images acquired using the conventional and the liquid-based cytology. Conventional cytology has overlapping cells, leukocytes, red blood cells, and mucus, and the image is hugely varied and significantly more complex for analysis than that of liquid-based cytology, which presents homogeneity in cell distribution and absence of obscuring factors[29].

Constructing and delivering a database of real images of conventional cytology from Pap smear will have considerable impact on computer vision and machine learning methods for knowledge discovery. It allows the comparison of existing algorithms, as well as the investigation of new prediction methodologies that are much more realistic from a biological perspective, and necessary to make such methods valuable to routine cervical cancer screening. Therefore, there are two major challenges: algorithms that allow the identification and classification of cells in a precise way, and, a complete, representative database with accurate information that enables the development/training of automatic methodologies to support the professional's decision.

This article presents a web platform, and uses this platform to support the development of a new cervical cell database, "CRIC Cervix", a collection of images obtained from conventional Pap smear. The Center for Recognition and Inspection of Cells ("CRIC") is a collaborative consortium among researchers that aims to provide cell collections to the scientific community. The CRIC Cervix images are similar to those obtained during an examination, with many cells per image and 150 dpi resolution. Cytopathologists classified the cells following the protocol defined by the Bethesda System nomenclature. In addition to the cells being individually identified and classified manually by different specialists, the full image field of collected cells also has a diagnosis. The information is available on the CRIC Searchable Image Database web platform, https://database.cric.com.br, which provides a publicly available web tool with information about cervical cells. The goal of this platform, as illustrated in Fig. 2, is to enable the community to explore cell morphology and variability, and facilitate discoveries from the image collections.

Results

Database contents. The collection CRIC Cervix has 400 images of Pap smears and 11,534 classified cells. In addition to https://database.cric.com.br, all the images in the current release are available in figshare[30] as a collection, https://doi.org/10.6084/m9.figshare.c.4960286.v2, under a Creative Commons Attribution 4.0 International license. The classification of the cells is provided in a CSV and JSON file formats with a companion README file that provides human readable details of CSV headers and JSON fields.

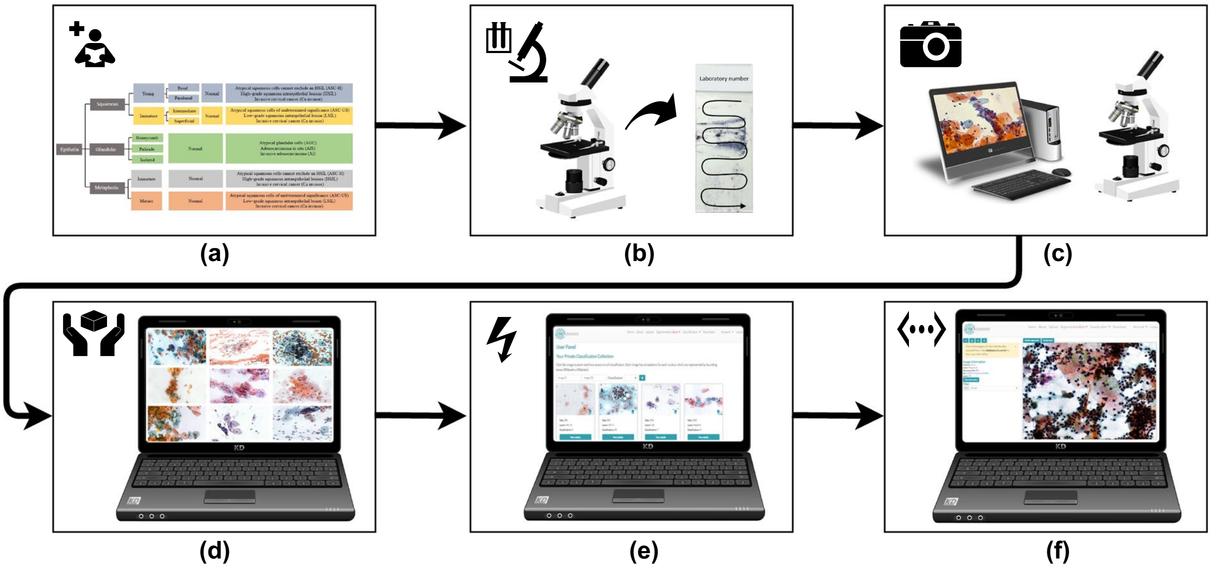

Fig. 2 CRIC Database workflow. (**a**) Latest protocols for taxonomy; (**b**) microscope screening for smear selection; (**c**) photo-documentation of smears; (**d**) image selection and curation; (**e**) insertion of images in CRIC, and (**f**) manual classification of cells by cytopathologists at CRIC.

Property	CRIC Cervix	Herlev	SIPaKMeD
Number of images	400	917	966
Cells per image	Variable	1	Variable
Image size (in pixels)	1,376 × 1,020	Variable	2,048 × 1,536
Resolution	0.228 μm/pixel	0.201 μm/pixel	Unknown
Classification	Manual	Manual	Manual
Classified cells	11,534	917	4,049
Validation	3 cytopathologists	2 cyto-technicians	expert cytopathologists
Download Page	database.cric.com.br	mde-lab.aegean.gr/downloads	www.cs.uoi.gr/~marina/sipakmed.html

Table 1. Comparison of properties among databases.

The collection is the first to cover conventional cytology cervical cells with classifications performed by several cytopathologists based on the Bethesda System nomenclature, which is the standardized terminology and most used worldwide in the area of cervical cytopathology. The cells in the CRIC Cervix collection are classified in six (6) classes: (1) negative for intraepithelial lesion or malignancy (NILM); (2) atypical squamous cells of undetermined significance, possibly non-neoplastic (ASC-US); (3) low-grade squamous intraepithelial lesion (LSIL); (4) atypical squamous cells, cannot exclude a high-grade lesion (ASC-H); (5) high-grade squamous intraepithelial lesion (HSIL); and (6) squamous cell carcinoma (SCC).

Another unique contribution of the proposed database is that it contains real images of conventional cytology with a variable number of cells, for example, including cells that overlap, which is often present in a smear of cytology, and must be considered when designing and testing automated cytology screening[23,31]. A list of differences among the cervical cell databases are listed in Tables 1 and 2.

Both in the total number of cells and the number of cells per class of lesion, CRIC Cervix is the largest (11,534) database available for real cancer cells, providing a broad diversity of neoplastic lesions to date. Also, the CRIC images contain many of challenges often found in ordinary exams, including overlapping cells and inflammatory cells. Classifications are performed by cytopathologists, using the standardized Bethesda System nomenclature. Although the Herlev database has more images, it contains far less cells. As for the Sipakmed database, it has isolated cells from Pap smear slides, and additionally uses a non-Bethesda-conforming classification, dividing cells into five categories: superficial-intermediate cells, parabasal cells, koilocytotic cells, dyskeratotic cells, and metaplastic cells, including normal, abnormal, and benign cases.

Web platform. In this work, we implemented the CRIC Searchable Image Database, available at https://database.cric.com.br, allowing the user to browse the images via a modern web-interface, easing data access and acceleration of discoveries (Supplementary Video). In addition to the web-interface, the platform is also accessible by a REST (Representational State Transfer) API that makes it possible to add interoperability with third-party projects. The CRIC Searchable Image Database backend uses several technologies, including the web framework Express[32] for Node.js[33], which supports the view system, routing, middleware, and other capabilities. In order to connect the web framework with the data, CRIC uses Sequelize[34] which offers Object Relational

	Cell count		
Cell type	**CRIC Cervix**	**Herlev**	**SIPaKMed**
NILM	6,779	144 (*)	(***)
ASC-US	606	0	(***)
ASC-H	925		(***)
LSIL	1,360	182	(***)
HSIL	1,703	493 (**)	(***)
SCC	161	0	(***)
Total	**11,534**	**819**	**4,049** (***)

Table 2. Number of cells per characteristics in databases. (*) Small requirements and Intermediate squamous epithelial. (**) Intermediate squamous epithelial and Severe squamous non-keratinizing dysplasia. (***) Cell categories cannot be translated into the Bethesda System nomenclature.

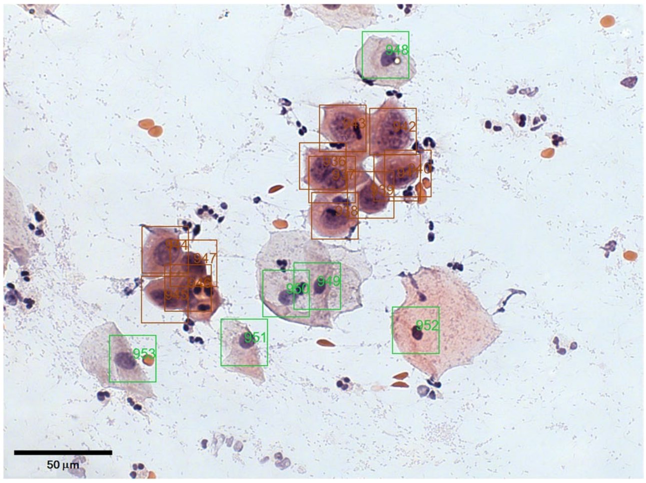

Fig. 3 CRIC Cervix Microscope Slide Image #383 with annotations.

Mapping (ORM) objects to the relational database systems, here enabled by MySQL[35]. The front-end uses Angular[36], which is a JavaScript framework for building web apps. The code is available at https://github.com/CRICDatabase/searchable-image-database and it is released under an open source license that allows anyone to reuse it, particularly if considering the documentation, which is available at https://cric-database.stoplight.io/docs/searchable-image-database-nodejs/reference/cric.v2.yaml.

After an image is uploaded to the CRIC Searchable Image Database, the user can identify and classify cells manually. The interface captures the coordinate of the click performed by the user, prompts the user for the cell classification, saves the information, and updates the image in the user's web browser with a square around the latest classified cell. The end result is a curated image such as Fig. 3. In addition to individual cell information, each image has a diagnosis following the most severe case among the classified cells in the image. The user can download the raw images, augmented images (such as Fig. 3), and associated data of the collection.

Figure 4 illustrates the platform dashboard where the user has an augmented view of the images with squares around the cells. If the user owns the image, they can add or edit the classification of the cells. We provided a playground area available at https://playground.database.cric.com.br/ where users may login to their own area to upload and classify their images. In the image upload, one must provide information about the captured data, author, a reference number from your laboratory, a DOI, if available, as seen at Fig. 4.

At the classification page, the user may access all the images of CRIC Cervix collection. They are presented in a reduced version with information on the slide, final diagnostic, and the number of cells in the image. When accessed in detail, the image is shown with the classification markings that may or may not be enabled. Lesions can also be accessed individually. The images can be saved, with or without the classification labels, and the labels can also be saved in JSON or CVS format. The files contain information on the lesion's classification and the position (x, y) in the image where it occurs. The markings were made in such a way as to be located inside the nucleus of the cell that contains the lesion.

At the download page, the user can save the entire collection, images, and classifications simultaneously or separately.

Discussion

We introduced the CRIC Searchable Image Database platform, which includes CRIC Cervix, a collection of cell images that are publicly available. This dataset of real images of conventional cervical cytology also have corresponding manual classification of pre-neoplastic cell lesions following the Bethesda System, and performed and certified by cytopathologists. The database will contribute significantly to the development of methodologies and/

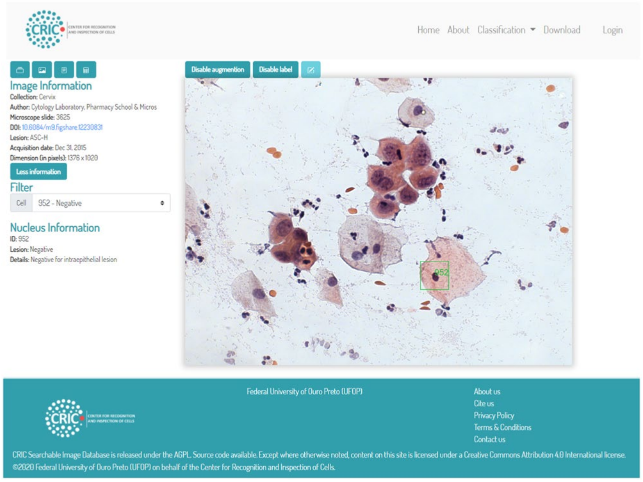

Image Information

Collection: Cervix
Author: Cytology Laboratory, Pharmacy School & Micros
Microscope slide: 3625
DOI: 10.6084/m9.figshare.12230831
Lesion: ASC-H
Acquisition date: Dec 31, 2015
Dimension (in pixels): 1376 x 1020

Filter

Cell: 952 - Negative

Nucleus Information

ID: 952
Lesion: Negative
Details: Negative for intraepithelial lesion

Federal University of Ouro Preto (UFOP)

CRIC Searchable Image Database is released under the AGPL. Source code available. Except where otherwise noted, content on this site is licensed under a Creative Commons Attribution 4.0 International license.
©2020 Federal University of Ouro Preto (UFOP) on behalf of the Center for Recognition and Inspection of Cells.

Fig. 4 CRIC Searchable Image Database currently available.

or automated products in computer-assisted cytology. It will also benefit researchers in machine learning who need labeled data for testing algorithm performance, particularly those related to deep convolutional neural networks[37].

The use of the CRIC platform and collections will allow discoveries regarding computational and biological descriptors that influence the decision about the presence of a lesion, supporting creation of new hypotheses about the differentiation of low and high grade cell lesions. We expect to further the development of image processing and computer analysis tools for the Pap smear, which are likely to augment the efficiency of visual inspection carried out by cytopathologists. Portraying the slide elements quantitatively allows the increase of the reliability and quality of the examination results. The goal is to reduce the false negatives, minimizing the number of lesions that are not detected or go unnoticed by professionals, providing additional information to support the decision of the specialist in cytopathology in order to obtain more accurate results.

CRIC Cervix has been used by the CRIC research group in several efforts. Araujo *et al.*[38] introduced a fast way for image ranking from large datasets using convolutional neural networks available in a tool named pyCBIR. They also developed a segmentation algorithm that applies a convolutional neural network trained with patch-images to identify and rank abnormal cell regions[39]. More recently, they presented an analysis framework, CRIC-feat, which streamlines the investigation of different image databases and respective descriptors, particularly applicable to Pap images[40]. Isidoro *et al.*[41], developed a methodology for automatic classification of cervical cell samples based on non-geometric characteristics present in cell nuclei, excluding the use of segmentation. Such studies can feasibly be explored in different ways by other researchers when the database becomes available, including new segmentation and cell classification methodologies.

An additional impact of the CRIC Searchable Image Database platform is an open-source web application for cell center annotation with single mouse-click and augmented view of the images with the cell annotation. The CRIC platform targets individual users or small groups without the financial resources to use commercially available digital histopathology software or the technical expertise to use complex and opaque solutions[42,43].

Besides commercially available digital histopathology software, e.g. provided by the manufacturers of the scanner hardware, there are many open source software solutions available for slide viewing and analysis, with some products supporting whole-slide images and others using standard graphics formats. Many solutions provide not only annotation capabilities, but also plug-in systems for automated analysis or pre-processing, e.g. Icy[44], CellProfiler[45] or SlideRunner[43]. However, none of the previous mentioned tools allow their users to publish the images to the general public.

The CRIC platform allows researchers to define collections, upload images they have, annotate cells in their images, share their images and associated annotations, and download the data to be used in computer vision and machine learning experiments.

At the moment, the CRIC Cervix collection includes normal squamous cells with precursor and invasive lesions. Updates will include normal glandular cells with precursor and invasive lesions. The authors opted for this strategy due to the fact that among invasive cervical cancers, the most common in the population is squamous cell carcinoma, which account for 75–90% of all cases, depending on the study[46–52].

Finally, the platform updates will increase the cytology collections and add more functionality at different user access levels, including manual segmentation of cervical cells. The CRIC Searchable Image platform also can support cytology research with other cell types, and we expect to include anal cytology soon.

Methods

The Research Ethics Committee approved this work at the Federal University of Ouro Preto, Minas Gerais, Brazil, through the document with protocol number 1944523. Due to limitations imposed by such a document, it is not possible to provide demographic information on the participants that contributed to the image dataset, including age and the ethnic diversity. This limitation exists because the participants did not provide consent for data sharing, and the requirement for informed consent was waived by the Research Ethics Committee under the agreements described above.

The Pap smear samples were obtained from female patients from the Southeast region of Brazil, South America. The Pap smear samples were processed and analyzed in the Cytology Laboratory of the Pharmacy School, Federal University of Ouro Preto, Minas Gerais, Brazil. Since the liquid-based cytology method is more expensive, and unavailable to this patient cohort, we employed the conventional cytology method for cytopathological smears, which is recommended by the Brazilian Ministry of Health. The conventional Pap test consists of introducing a speculum into the vaginal canal for exposure and inspection of the uterine cervix. The Ayre spatula collects cells from the ectocervix and the endocervical brush from the endocervix. The collected material was dispensed on a matte tip blade, previously identified in pencil with the patient's initials and date of birth. In order to preserve the cell structure, the material was fixed in 96% alcohol, and placed in slide bottles supplied by the Cytology Laboratory. These slide bottles were placed in plastic bags, sealed, and transported to the Cytology Laboratory for the processing, and analysis.

In the Cytology Laboratory, a pre-analytical evaluation was carried out, which consisted of selecting the smears that were in adequate conditions for the analysis, and were duly identified with the laboratory registration number. Papanicolaou staining was used, consisting of a nuclear dye, hematoxylin, and two cytoplasmic stains, Orange G and EA 36 or EA 50. Leica AutoStainer XL was used for automatic staining. Then, the smears were assembled automatically, with Entellan between the slide and the coverslip, using the Leica CV5030 equipment. The microscopic analysis, which includes the observation and evaluation of cervical cells under an optical microscope to classify cervical cancer precursor lesions, was performed by a team of 3 cytopathologists. Since 2013, these professionals conduct cytopathological exams for the Brazilian Unified Health System, also known as SUS, and since 2014, they have been carrying out External Quality Monitoring of laboratories that provide services to the SUS in the state of Minas Gerais, Brazil. The entire process was carried out following the Brazilian Ministry of Health recommendation[53].

The Cytology Laboratory performs Internal Quality Monitoring, which are regular actions to guarantee the quality of the Pap smear, including control of the receipt, assembly and coloration of the samples, use of the 100% Rapid Review method (negative smears are reviewed) and Rapid Pre-scrutiny (scrutiny prior to routine reading), implementation of corrective actions, recording of results, promotion of permanent education to professionals and internal audit[54].

The first step in the development of the CRIC web platform was the survey of the most recent protocols for taxonomy and the elaboration of a diagram with cell types and respective lesions for guidance in the selection of Pap smears, which served as an interface model used to create the decision tree on the platform.

The smears selected for photo-documentation come from the collection of cervical-vaginal smears obtained through conventional cytology performed at the Cytology Laboratory as part of the routine care. The collection was curated by analyzing the database to access the results of the exam. From the result, the smears were scrutinized under an optical microscope by the three cytologists in order to assess the cytomorphological criteria that best represented each cytological alteration. After consensus among the three specialists, 118 smears from 118 patients were selected. These 118 patients are diagnosed with 3 NILM, 45 ASC-US, 45 LSIL, 12 ASC-H, 10 HSIL and 3 SCC.

The smear photo-documentation was performed by conventional bright field microscopy with a 40× objective and a 10× eyepiece, using a Zeiss AxionCam MRc digital camera coupled to the Zeiss AxioImager.Z2 microscope, with the Axion Vision Zeiss Software, which are at the Multi-User Laboratory of the Biological Sciences Research Center (NUPEB), Federal University of Ouro Preto, Minas Gerais, Brazil. After the photo-documentation, all images obtained by the three cytologists were analyzed and curated, and 400 images were selected to compose the CRIC Cervix dataset.

The classification of cervical precursor and invasive lesions was performed within the CRIC Searchable Image platform, https://database.cric.com.br, according to the Brazilian Nomenclature for cervical cytopathological reports, based on the Bethesda System nomenclature[27]. The classification was carried out by three members of the Cytology Laboratory team, cytologists with experience of 6, 11 and 20 years working in diagnostic cytology.

Each cell was classified by selecting the class corresponding to the lesion, based on morphological criteria prescribed by the Bethesda System nomenclature, followed by marking the center of the cell nucleus. This procedure repeated until all the cells in the image were labeled. The classification protocol followed by the three cytologists started with an independent classification made by an initial professional, then the second specialist checked the labels. Next, the third cytologist performed a review of the markings and if the three answers were in agreement, he/she approved the label. Otherwise, a consensus was reached among the three cytologists to define the final label. After manual labeling, computational routines for pre-processing removed labeled cells near the borders and whose nucleus or cytoplasm structures were not fully visible. This process aims to avoid feeding algorithms with incorrect information about the nucleus or cytoplasm.

Data availability

The collection CRIC Cervix data described in this paper is available at https://database.cric.com.br, and also at https://doi.org/10.6084/m9.figshare.c.4960286.v2[30].

Code availability

The classifications where made using https://database.cric.com.br. The source code of the website is available on GitHub, https://github.com/CRICDatabase/searchable-image-database, under GNU Affero General Public License v3.0.

Received: 2 December 2020; Accepted: 11 May 2021;
Published online: 10 June 2021

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA: a cancer journal for clinicians* **67**, 7–30, https://doi.org/10.3322/caac.21387 (2017).
2. World Health Organization. International Agency for Research on Cancer. *Globocan 2020 - The Global Cancer Observatory*, http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx (2021).
3. Meggiolaro, A. *et al*. The role of pap test screening against cervical cancer: a systematic review and meta-analysis. *La Clinica terapeutica* **167**, 124–139, https://doi.org/10.7417/CT.2016.1942 (2016).
4. Sachan, P. L., Singh, M., Patel, M. L. & Sachan, R. A study on cervical cancer screening using pap smear test and clinical correlation. *Asia-Pacific journal of oncology nursing* **5**, 337–341, https://doi.org/10.4103/apjon.apjon_15_18 (2018).
5. Zhao, X. *et al*. The performance of human papillomavirus dna detection with type 16/18 genotyping by hybrid capture in primary test of cervical cancer screening: a cross-sectional study in 10669 chinese women. *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **24**, 1322–1327, https://doi.org/10.1016/j.cmi.2018.02.027 (2018).
6. Saslow, D. *et al*. American cancer society, american society for colposcopy and cervical pathology, and american society for clinical pathology screening guidelines for the prevention and early detection of cervical câncer. *CA: a cancer journal for clinicians* **62**, 147–172, https://doi.org/10.3322/caac.21139 (2012).
7. US Preventive Services Task Force. *et al*. Screening for cervical cancer: Us preventive services task force recommendation statement. *JAMA* **320**, 674–686, https://doi.org/10.1001/jama.2018.10897 (2018).
8. Ferlay, J. *et al*. Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018. *European journal of cancer* **103**, 356–387, https://doi.org/10.1016/j.ejca.2018.07.005 (2018).
9. do Nascimento Tavares, S. B. *et al*. Internal quality control for cervical cytopathology: Comparison of potential false-negatives detected at rapid prescreening and at 100% rapid review. *Acta cytologica* **58**, 439–445, https://doi.org/10.1159/000368041 (2014).
10. Ramalho, G. L. B. *et al*. Cell reconstruction under Voronoi and enclosing ellipses from 3d microscopy. *IEEE International Symposium on Biomedical Imaging (ISBI)* (2015).
11. Rezende, M. T., Bianchi, A. G. & Carneiro, C. M. Cervical cancer: Automation of pap test screening. *Diagnostic Cytopathology* https://doi.org/10.1002/dc.24708 (2021).
12. Pycock, D. & Taylor, C. J. The magiscan image analyser as a diagnostic aid in cytology. *Analytical and quantitative cytology* **3**, 49–54 (1981).
13. Taylor, R. N. *et al*. Cytoview: A prototype computer image-based papanicolaou smear proficiency test. *Acta cytologica* **43**, 1045–1051, https://doi.org/10.1159/000331352 (1999).
14. Mahovlić, V. *et al*. Digital morphometry of cytologic aspirate endometrial samples. *Collegium antropologicum* **34**, 45–51 (2010).
15. Lepe, M., Eklund, C. M., Quddus, M. R. & Paquette, C. Atypical glandular cells: interobserver variability according to clinical management. *Acta cytologica* **62**, 397–404, https://doi.org/10.1159/000489968 (2018).
16. Wilkinson, M. D. *et al*. The fair guiding principles for scientific data management and stewardship. *Scientific data* **3**, 1–9 (2016).
17. Jantzen, J., Norup, J., Dounias, G. & Bjerregaard, B. Pap-smear benchmark data for pattern classification. In NiSIS (ed.) *NiSIS 2005: Nature inspired Smart Information Systems (NiSIS), EU co-ordination action*, 1–9 (2005).
18. Chankong, T., Theera-Umpon, N. & Auephanwiriyakul, S. Automatic cervical cell segmentation and classification in pap smears. *Computer methods and programs in biomedicine* **113**, 539–556, https://doi.org/10.1016/j.cmpb.2013.12.012 (2014).
19. Singh, U. *et al*. Comparative study between liquid-based cytology & conventional pap smear for cytological follow up of treated patients of cancer cervix. *The Indian journal of medical research* **147**, 263–267, https://doi.org/10.4103/ijmr.IJMR_854_16 (2018).
20. Lu, Z. *et al*. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE Journal of Biomedical and Health Informatics* **21**, 441–450, https://doi.org/10.1109/JBHI.2016.2519686 (2016).
21. Bora, K., Chowdhury, M., Mahanta, L. B., Kundu, M. K. & Das, A. K. Automated classification of pap smear images to detect cervical dysplasia. *Computer Methods and Programs in Biomedicine* **138**, 31–47, https://doi.org/10.1016/j.cmpb.2016.10.001 (2017).
22. Zhang, L. *et al*. Graph-based segmentation of abnormal nuclei in cervical cytology. *Computerized Medical Imaging and Graphics* **56**, 38–48, https://doi.org/10.1016/j.compmedimag.2017.01.002 (2017).
23. William, W., Ware, A., Basaza-Ejiri, A. H. & Obungoloch, J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. *Computer methods and programs in biomedicine* **164**, 15–22, https://doi.org/10.1016/j.cmpb.2018.05.034 (2018).
24. Lakshmi, G. K. & Krishnaveni, K. Feature extraction and feature set selection for cervical cancer diagnosis. *Indian Journal of Science and Technology* **9**, 1–7, https://doi.org/10.17485/ijst/2016/v9i19/93881 (2016).
25. Plissiti, M. E. *et al*. Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 3144–3148 (2018).
26. Wilbur, D. C. & Nayar, R. Bethesda 2014: improving on a paradigm shift. *Cytopathology: official journal of the British Society for Clinical Cytology* **26**, 339–342, https://doi.org/10.1111/cyt.12300 (2015).
27. Nayar, R. & Wilbur, D. C. *The Bethesda system for reporting cervical cytology: definitions, criteria, and explanatory notes* (Springer, 2015).
28. Ushizima, D., Bianchi, A. & Carneiro, C. Segmentation of subcellular compartments combining superpixel representation with Voronoi diagrams. *IEEE International Symposium on Biomedical Imaging (ISBI)* (2014).
29. Phaliwong, P. *et al*. Cervical cancer detection between conventional and liquid based cervical cytology: a 6-year experience in northern bangkok thailand. *Asian Pacific journal of cancer prevention* **19**, 1331–1336, https://doi.org/10.22034/APJCP.2018.19.5.1331 (2018).
30. Rezende, M. T. *et al*. Cric cervix cell classification. *figshare* https://doi.org/10.6084/m9.figshare.c.4960286.v2 (2020).
31. Thrall, M. J. Automated screening of papanicolaou tests: A review of the literature. *Diagnostic Cytopathology* **47**, 20–27, https://doi.org/10.1002/dc.23931 (2019).
32. OpenJS Foundation. *Express 4.x - API Reference*, https://expressjs.com/pt-br/api.html (2017).
33. OpenJS Foundation. *Node.js v14.8.0 Documentation*, https://nodejs.org/en/docs/ (2020).
34. Sascha Depold and others. *Sequelize v6 API Reference*, https://sequelize.org/ (2014).
35. Oracle Corporation. *MySQL 8.0 Reference Manual*, https://dev.mysql.com/doc/refman/8.0/en/ (2020).
36. Jain, N., Bhansali, A. & Mehta, D. Angularjs: A modern mvc framework in javascript. *Journal of Global Research in Computer Science* **5**, 17–23 (2014).
37. Ushizima, D. *et al*. IDEAL: Images across Domains, Experiments, Algorithms and Learning. *Journal of Minerals, Metals and Materials* https://doi.org/10.1007/s11837-016-2098-4 (2016).
38. Araujo, F. H. D. *et al*. Reverse image search for scientific data within and beyond the visible spectrum. *Expert Systems with Applications* **109**, 35–48, https://doi.org/10.1016/j.eswa.2018.05.015 (2018).
39. Araújo, F. H. D. *et al*. Deep learning for cell image segmentation and ranking. *Computerized Medical Imaging and Graphics* **72**, 13–21, https://doi.org/10.1016/j.compmedimag.2019.01.003 (2019).
40. Araújo, F. H. D. *et al*. Active contours for overlapping cervical cell segmentation. *Int. J. Biomedical Engineering and Technology* **35**, 70–92, https://doi.org/10.1504/IJBET.2021.112834 (2021).

41. Isidoro, D. *et al*. Automatic classification of cervical cell patches based on non-geometric characteristics: In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 845–852, https://doi.org/10.5220/0009172208450852 (SCITEPRESS - Science and Technology Publications, 2020).
42. Bertram, C. A. & Klopfleisch, R. The pathologist 2.0: An update on digital pathology in veterinary medicine. *Veterinary pathology* **54**, 756–766, https://doi.org/10.1177/0300985817709888 (2017).
43. Aubreville, M., Bertram, C., Klopfleisch, R. & Maier, A. Sliderunner: A tool for massive cell annotations in whole slide images. *Informatik aktuell* 309–314, https://doi.org/10.1007/978-3-662-56537-7_81 (2018).
44. De Chaumont, F. *et al*. Icy: an open bioimage informatics platform for extended reproducible research. *Nature methods* **9**, 690–696, https://doi.org/10.1038/nmeth.2075 (2012).
45. Jones, T. R. *et al*. Cellprofiler analyst: data exploration and analysis software for complex image-based screens. *BMC bioinformatics* **9**, 1–16, https://doi.org/10.1186/1471-2105-9-482 (2008).
46. Vizcaino, A. P. *et al*. International trends in the incidence of cervical cancer: I. adenocarcinoma and adenosquamous cell carcinomas. *International journal of cancer* **75** 536–545, https://doi.org/10.1002/(sici)1097-0215(19980209)75:4<536::aid-ijc8>3.0.co;2-u (1998).
47. Vizcaino, A. P. *et al*. International trends in incidence of cervical cancer: Ii. squamous-cell carcinoma. *International journal of cancer* **86**,429–435, https://doi.org/10.1002/(sici)1097-0215(20000501)86:3<429::aid-ijc20>3.0.co;2-d (2000).
48. Parkin, D. M. & Bray, F. Chapter 2: The burden of hpv-related cancers. *Vaccine* **24**, S3/11–25, https://doi.org/10.1016/j.vaccine.2006.05.111 (2006).
49. Bosch, F. X. *et al*. Epidemiology and natural history of human papillomavirus infections and type-specific implications in cervical neoplasia. *Vaccine* **26**, K1–16, https://doi.org/10.1016/j.vaccine.2008.05.064 (2008).
50. Tjalma, W. A. *et al*. Differences in human papillomavirus type distribution in high-grade cervical intraepithelial neoplasia and invasive cervical cancer in europe. *International journal of cancer* **132**, 854–867, https://doi.org/10.1002/ijc.27713 (2013).
51. Quek, S. C. *et al*. Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical intraepithelial neoplasia across 5 countries in asia. *International journal of gynecological cancer: official journal of the International Gynecological Cancer Society* **23**, 148–156, https://doi.org/10.1097/IGC.0b013e31827670fd (2013).
52. Castanon, A., Landy, R. & Sasieni, P. D. Is cervical screening preventing adenocarcinoma and adenosquamous carcinoma of the cervix? *International journal of cancer* **139**, 1040–1045, https://doi.org/10.1002/ijc.30152 (2016).
53. Instituto Nacional de Câncer José Alencar Gomes da Silva. Manual of Quality Management for Cytopathology Laboratories (Ministério da Saúde, 2016).
54. Branca, M. & Longatto-Filho, A. Recommendations on quality control and quality assurance in cervical cytology. *Acta cytologica* **59**, 361–369, https://doi.org/10.1159/000441515 (2015).

Acknowledgements

This work was supported by CAPES/CNPq-PVE (401442/2014-4), CNPq Fellowship (308947/2020-7, 305895/2019-2, 304673/2011-0, 306396/2015-7), CNPq (444784/2014-4), CNPq Rebrats (401120/2013-9), CNPq, PPSUS/FAPEMIG (APQ-03740-17), FAPEMIG (APQ-02369-14), FAPEMIG (APQ-00751-19), PROPPI/UFOP, UFOP (23109.000929/2020-88, 23109.000928/2020-33, 23109.003209/2016-98, 23109.003515/2018-96 and 23109.003517/2018-85), and Office of Science, U.S. Department of Energy (DOE) under Contract No. DE-AC02-05CH11231. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 (88882.459660/2019-01). The authors acknowledge the Laboratório Multiusuários de Microscopia Avançada e Microanálise, Núcleo de Pesquisas em Ciências Biológicas, Universidade Federal de Ouro Preto, for providing the equipment and technical support for experiments involving brightfield.

Author contributions

Silva led the final development of https://database.cric.com.br, and its validation, that started with Oliveira, Bernardo, Machado, and Costa. Rezende photodocumented the smears to obtain the images and led the cell classification that was reviewed by Carneiro and Tobias. Carneiro, Bianchi, Ushizima, and Medeiros formulated the project, and designed algorithms. Carneiro and Bianchi are responsible for the leadership in this research activity planning, execution, and coordination. Rezende, Silva, Bianchi, Carneiro, Ushizima, and Medeiros contributed to writing the paper.

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/s41597-021-00933-8.

Correspondence and requests for materials should be addressed to M.T.R. or 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.