RIVA: An Image Dataset of Conventional Pap Smear Cytology with Multiple Independent Annotations

Autoría no Ditelliana: Pérez Bianchi, Paula; Anselmo, Sol; Vásquez Currié, Malena; Medel, Jimena; Uelf, Estefanía; Dos Santos, Alicia; Buosi, Noemí; Vargas, Rosana; Reves Szemere, Juliana; Volcovinsky, Bruno; Massaroli, Hugo; Monastra, Alejandro; Bruno, Luciana

Autorías Ditellianas: Andrade, Manuel; Iarussi, Emmanuel; Siless, Viviana

Fecha de publicación 09/12/2025

Publicado originalmente en: Scientific Data (e-ISSN: 2052-4463)

¿Cómo citar este trabajo?

Pérez Bianchi, P., Anselmo, S., Vásquez Currié, M. et al. RIVA: An Image Dataset of Conventional Pap Smear Cytology with Multiple Independent Annotations. Sci Data 12, 1991 (2025). https://doi.org/10.1038/s41597-025-06280-2

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

RIVA: An Image Dataset of Conventional Pap Smear Cytology with Multiple Independent Annotations

Paula Perez Bianchi[1], Sol Anselmo[2,10], Malena Vásquez Currié[2,10], Jimena Medel[3], Estefanía Uelf[3], Alicia Dos Santos[3], Noemí Buosi[3], Rosana Vargas[3,4], Juliana Reves Szemere[5], Bruno Volcovinsky[2], Hugo Massaroli[1], Manuel Andrade[6], Alejandro Monastra[7,8], Emmanuel Iarussi[6,8], Viviana Siless[4,6] & Luciana Bruno[8,9] ✉

The Pap smear remains the primary screening test for cervical cancer in many low-resource regions, yet publicly available image datasets largely feature liquid-based preparations. We introduce RIVA, a high-resolution collection of 959 conventional-smear images (1024 × 1024 px) scanned at 40x magnification, sourced from 115 patients. To ensure label quality, each image was annotated by up to four independent medical professionals, with 42% of the images reviewed by all four, resulting in 26,158 annotations based on the Bethesda classification. Annotations provide coordinates of nuclei and classification labels by up to four annotators. The dataset includes 15,949 unique cells across five (pre) cancerous types (SCC, HSIL, ASCH, LSIL, ASCUS) and three non-lesion categories (NILM, ENDO, INFL). These four-expert annotations not only give RIVA a consensus-driven ground truth for robust AI training but also enable inter-annotator consistency analysis-agreement rates reach 94% for lesion vs. non-lesion and 74% across the full eight-category Bethesda scheme.

Background & Summary

Cervical cancer is one of the most preventable forms of cancer, yet it remains a significant public health challenge worldwide, particularly in low- and middle-income countries (LMICs) where access to screening programs is limited. Globally, over 600,000 new cases and more than 340,000 deaths were reported in 2020 alone, with nearly 90% of these deaths occurring in LMICs[1]. Early detection and classification of cervical lesions are crucial, as they significantly increase the chances of successful treatment and long-term survival.

Papanicolaou (Pap) smear screening has proven highly effective in reducing the incidence and mortality associated with cervical cancer[2]. However, conventional cytology requires manual inspection of stained cell samples-a time-consuming and expertise-dependent process that is susceptible to human error, inter-observer variability, and reduced sensitivity and reproducibility, especially under high workloads[3]. These limitations pose a serious barrier to widespread and equitable implementation, particularly in settings with constrained healthcare resources.

Recent advances in artificial intelligence (AI), particularly deep learning, have enabled the automation of numerous medical imaging tasks, achieving or even surpassing expert-level performance in certain domains[4]. Despite these advancements, the application of AI to cytology remains relatively underdeveloped, and its clinical

[1]Departamento de Computación (DC), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina. [2]Instituto Tecnológico de Buenos Aires (ITBA), Buenos Aires, Argentina. [3]Servicio de Anatomía Patológica, Hospital Bernardino Rivadavia, Buenos Aires, Argentina. [4]Pathology & Molecular Biology Laboratories, Biomakers, Buenos Aires, Argentina. [5]Escuela de Ciencia y Tecnología, Universidad Nacional de San Martín, Buenos Aires, Argentina. [6]Universidad Torcuato Di Tella (UTDT), Buenos Aires, Argentina. [7]Instituto de Ciencias, Universidad Nacional de General Sarmiento, Buenos Aires, Argentina. [8]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires, Argentina. [9]Instituto de Cálculo (IC), Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires (UBA), Buenos Aires, Argentina. [10]These authors contributed equally: Sol Anselmo, Malena Vásquez Currié. ✉e-mail: luciana.bruno@ic.fcen.uba.ar

Features	Herlev[7]	Sipakmed[8]	CRIC[9]	APACC[10]	RIVA (Ours)[13]
Number of smears/patients	—	—	118	107	115
Number of images	917	966	400	21,371	959
Number of cells	917	4,049	11,534	103,675	15,949
Image size (in pixels)	Variable	2,048 × 1,536	1,376 × 1,020	2,000 × 2,000	1,024 x 1,024
Number of classes	7	5	6	4	8
TBS*	No	No	Yes	No	Yes
Annotators	2 cyto-technicians (+1 doctor)	expert cytopathologists	3 cytopathologists	3 cytopathologists	4 cytopathologists
Multiple annotations/cell**	—	—	No	No	Yes

Table 1. Comparison of Pap smear cytology image datasets. TBS* indicates alignment with the Bethesda System for cell classification[16]. **30% of the cells were independently annotated by more than one pathologist.

deployment is rare. A major bottleneck is the lack of large, high-quality, and diverse datasets necessary for training robust and generalizable models[5,6]. A review of existing datasets reveals four publicly available collections: Herlev[7], SIPaKMeD[8], CRIC Cervix[9], and APACC[10] (Table 1).

The Herlev[7] database is a well-known benchmark for cervical cytology analysis that contains a total of 917 single-cell images, each depicting an isolated cervical cell. These images are classified into seven diagnostic categories representing different stages of pre-neoplastic lesions. The Herlev dataset has been widely adopted in early studies applying machine learning to cervical cell classification due to its high-quality annotations and clear class definitions. Nevertheless, its limited sample size and deviation from the Bethesda System reduce model generalizability and clinical compatibility.

The SIPaKMeD[8] dataset is a publicly available resource consisting of 966 image patches and 4,049 isolated cervical cell images extracted from Pap smear slides. Each image is annotated with both segmentation and cell category, which falls into one of five classes. The images are captured using a CCD camera (Infinity 1 Lumenera) mounted on an OLYMPUS BX53F optical microscope. One of the key strengths of SIPaKMeD is the availability of both classification and segmentation labels, making it a valuable resource for multi-task learning applications. Nevertheless, its limited size and the absence of Bethesda-based classification reduce its suitability for deep learning approaches and clinical diagnostic use.

The CRIC[9] (Center for Recognition and Inspection of Cells) database comprises 11,534 annotated cell images obtained from 118 patients, classified into six distinct categories following the Bethesda System. A major strength of this dataset is its adherence to clinically relevant classification standards, with annotations performed by multiple expert cytopathologists. It further distinguishes itself by offering high magnification (40x), which captures fine cellular detail, and by providing a user-friendly online interface that facilitates dataset exploration. One limitation, however, is that the cells were manually selected, which may introduce sampling bias by potentially favoring more visually distinctive or diagnostically relevant cells over the true cell distribution on whole slides. In addition, images were acquired using conventional microscopy under variable illumination settings that are difficult to reproduce consistently across different acquisition systems, may hinder reproducibility in AI-based applications.

The APACC[10] (Annotated PAp cell images and smear slices for Cell Classification) database contains 103,675 annotated cell images extracted from 107 whole-slide Pap smear samples. To support more localized analysis, these samples are further divided into 21,371 sub-regions. A major strength of APACC is the use of a commercial scanner, which ensures image quality reproducibility. The dataset also benefits from a large number of labeled cells and the random selection of image patches, making the data more representative of whole-slide distributions. Despite these advantages, APACC includes annotations for only four cell classes that do not fully align with the Bethesda system.

Labeling strategies in cervical cytology databases vary widely, influencing label quality and comparability[11,12]. For instance, CRIC employs simultaneous double annotation with adjudication, while APACC uses independent annotations with expert review only when discrepancies arise. In this paper, we introduce RIVA[13]-named after Hospital Rivadavia, where the samples were collected-a database of digitized conventional Pap smear cytology images independently annotated by multiple experts. Each cell in RIVA is labeled by up to four expert pathologists according to the Bethesda System, with all individual annotations made available. Among all the cells, 30% were annotated by more than one expert. This design enables detailed inter-annotator agreement analysis and consensus studies, offering a robust foundation for developing and evaluating AI-based cytology tools.

The RIVA[13] dataset comprises 959 image mini-patches of 1024 × 1024 pixels, extracted from whole-slide scans of samples from 115 patients using a Grundium Ocus 40 scanner. Patch sampling was automated to reduce selection bias. A subset of 400 out of the 959 image patches was independently annotated by four experts, while the remaining patches were annotated by a single pathologist. Although only a portion of the dataset received multiple annotations, this subset represents a substantial advance over existing datasets. In total, RIVA includes 26,158 annotations corresponding to 15,949 unique cells, categorized into five (pre) cancerous types-SCC, HSIL, ASCH, LSIL, ASCUS-and three non-lesion categories-INFL (inflammatory cells), ENDO (endocervical cells), and NILM (negative for intraepithelial lesion or malignancy). Each annotation provides the approximate nucleus center and a cell type label, as independently assigned by one to four expert pathologists. To ensure broad accessibility, we developed a publicly available web (https://beta-digitalpapsdb.exactas.uba.ar/) that enables users to explore the Pap smear images, review expert annotations, and download the dataset.

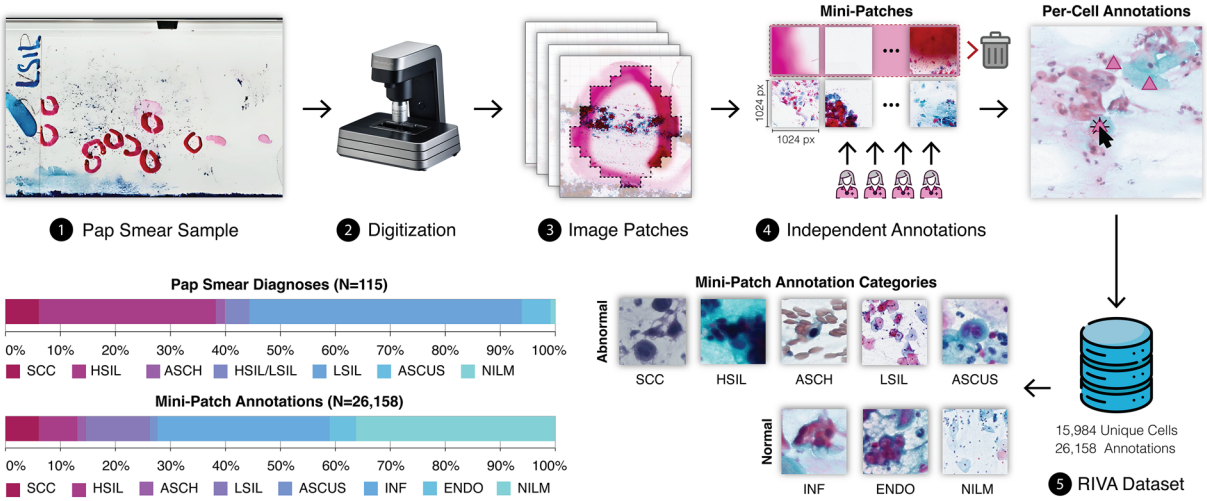

Fig. 1 Construction Pipeline. **(1)(2)** Whole-slide Pap smear samples with diagnostic markings were scanned using a high-resolution slide scanner. **(3)** Only marked diagnostic regions of interest were selected and digitized. **(4)** From these digitized areas, 1024 × 1024 pixel image mini-patches were extracted. Independent expert annotators reviewed the patches, discarding irrelevant ones and providing per-cell annotations on the selected mini-patches. **(5)** The resulting dataset, named RIVA, contains 15,949 unique cells and 26,158 total annotations. The annotations are categorized into abnormal classes-SCC, HSIL, ASCH, LSIL, ASCUS-and normal classes-INF, ENDO and NILM. The bar plots below summarize the distribution of categories at two levels: whole-slide diagnoses (N = 115 samples) and mini-patch annotations (N = 26,158), showing a greater class diversity at the patch level due to the presence of heterogeneous cell populations within slides.

Methods

Sample selection. Pap smear samples were obtained from the permanent repository of the Pathology Laboratory at Hospital Bernardino Rivadavia, Buenos Aires. They were collected from female patients in Buenos Aires and surrounding areas. The study protocol was approved by the hospital authorities (see Ethics Information). Pathologists at the hospital selected samples for scanning based on two criteria: a confirmed positive diagnosis for lesions and good sample condition. Additionally, lesion-free samples were included solely based on their quality. No other selection criteria were applied In total, the dataset comprises samples from 115 patients, distributed across the following diagnostic categories: SCC (7), HSIL (37), ASCH (2), HSIL/LSIL (5), LSIL (57), ASCUS (6), and NILM (1) (see Fig. 1). Note that some samples exhibited HSIL and LSIL lesions simultaneously. To ensure ethical compliance, all samples were anonymized and no cross-referenced patient data was used.

Digitalization. Prior to scanning, samples were inspected by both a technician and a pathologist, and restored when necessary to ensure optimal image quality. Slides were digitized using a Grundium Ocus 40 scanner (https://www.grundium.com/scanners/ocus40/) at 40x magnification, achieving a resolution of 0.25 μm/pixel as specified by the manufacturer. To accelerate the process, limit file sizes, and ensure sample diversity, only specific regions of the slides were scanned. These regions, previously identified by pathologists and marked with a permanent marker, contained both lesion-bearing and normal cells (see Fig. 1). Each sample required approximately 10 minutes to scan, and the process was carried out by experts.

Mini-patch Generation. Patches were manually extracted from the samples using Aperio Image Scope to create smaller SVS files compatible with the Python processing pipeline. These patches were then converted into `.png` format using a custom algorithm, as Label Studio[14]—the annotation software used by medical professionals-does not support the `.svs` format. Subsequently, we developed a script to subdivide each extracted patch into uniform non-overlapping mini-patches of 1024 × 1024 pixels (see Fig. 1).

During the selection process, priority was given to mini-patches that enabled medical professionals to make the greatest number of annotations. As a quality criterion, we focused on patches containing clearly identifiable cells, deliberately excluding background regions. To automate the initial selection, we applied an algorithm based on the Otsu method, using its threshold value to estimate the proportion of white pixels and identify likely background areas. This process resulted in a set of 1,000 mini-patches for annotation.

Annotations. We used Label Studio[14] as the annotation platform, selecting the *Keypoint Labeling* interface, which allowed medical professionals to annotate by clicking once on the category and once on the cell's nucleus. Label Studio was run on a local installation to ensure data privacy and control throughout the annotation process. After preliminary testing, we developed a detailed instruction manual and recorded an instructional video to guide the annotation process. To ensure reliable data for assessing inter-annotator agreement, each cytopathologist was assigned a different project inside the tool, allowing independent annotations for every mini-patch. Patches were presented in random order, without any indication of their origin or grouping, to minimize potential bias during annotation. In the first batch, 400 mini-patches were uploaded to each project to identify and

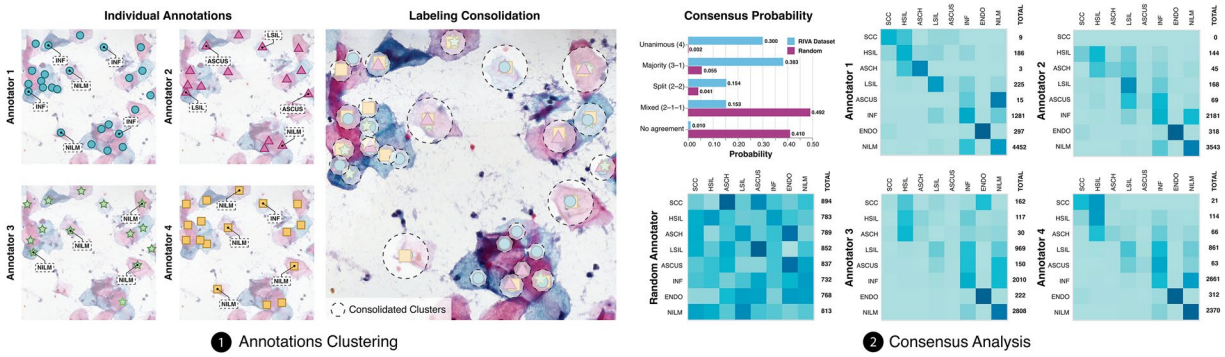

Fig. 2 Annotations Clustering **(1)**. Individual annotations from four pathologists on the same image. Each marker represents a cell labeled by a specific annotator. Consolidated clusters obtained by grouping annotations referring to the same cell using the MeanShift algorithm. Dashed circles indicate the resulting consensus cell positions. Consensus Analysis **(2)**. Top left: Distribution of consensus, comparing the empirical data (RIVA Dataset) to a random baseline. Unanimous agreement occurs in 30% of cases, while full disagreement is rare. The remaining charts display confusion matrices for each annotator (rows: annotator labels, columns: other 3 annotators labels), along with a random baseline, illustrating that most disagreements occur between neighboring or related diagnostic classes.

analyze annotation discrepancies. The remaining 600 mini-patches were distributed evenly, with 150 patches assigned to each doctor, significantly expanding our database. Annotators were encouraged to tag all recognizable cells within the mini-patches. Annotators performed their tasks independently and were not exposed to the annotations made by others, ensuring that each image was labeled without bias from prior opinions. Following the initial annotation round, we conducted a manual curation process to clean the data and identify issues. This revealed cases where images were overlooked or where annotations were discarded due to the absence of identifiable cells. These cases were resolved by asking the cytopathologists to re-annotate the affected images. The final curated dataset consisted of 959 annotated mini-patches.

Procedure for Clustering Annotations. Before analyzing annotation agreement, we first needed to group nearby annotations that likely referred to the same cell. To achieve this, we used the *MeanShift* clustering algorithm[15] (see Fig. 2. Because clustering performance depends on the choice of a bandwidth parameter, we introduced a consistency rule: no cluster could contain more than one annotation from the same pathologist. This ensured that each cluster represented a unique cell rather than overlapping or ambiguous markings. In cases where a cluster violated this rule-suggesting that two nearby cells may have been merged-we re-applied *MeanShift* using a smaller bandwidth, allowing the cluster to be split into distinct cell detections. Since the annotations correspond to cell nuclei, we selected the final coordinates for each cell as the most centrally located annotation within its cluster-essentially the one that best represents the average or consensus location of that cell. This guarantees that all reported positions in the curated dataset directly match a point marked by a pathologist. Certain edge cases required additional attention, such as images with high cell density (e.g., overlapping tissue flaps) or LSIL cells that may present with two nuclei. These were carefully reviewed to ensure the integrity of the resulting dataset.

Probabilistic Model for Inter-Annotator Agreement. To quantify the consistency and reliability of expert annotations, we introduce a probabilistic model that relates annotation agreement to the underlying confidence of classification decisions. Thus, we consider the pathology level of a cell as a continuous variable, x_0, and define $p_i(x_0)$ as the probability that an annotator classifies this case into class i, where $1 \leq i \leq M$, with M being the total number of classes. For random annotators, $p_i = 1/M$ for all i, whereas for skilled annotators, the probabilities will peak at a particular index j, with $p_j = p_{\max}$ approaching one and the remaining p_i values being close to zero. Given N annotators of comparable skill, and a cell with pathology x_0, the model enables the computation of probabilities for different agreement scenarios. For instance, the probability that all annotators assign the same class to a given cell, i.e. unanimous agreement, is defined as:

$$\mathscr{P}_N = \sum_{i=1}^{M} (p_i)^N. \tag{1}$$

Similarly, the probability that exactly $N-1$ annotators agree on the same class, while one annotator assigns a different classification, i.e. majority agreement, is:

$$\mathscr{P}_{N-1} = N \sum_{i=1}^{M} (p_i)^{N-1}(1 - p_i). \tag{2}$$

Equivalent expressions can be derived for other agreement scenarios. In cases where a single probability $p_{\max}$ dominates (i.e., is significantly higher than the others) and the number of annotators is large, these sums are

primarily governed by p_{max}, while contributions from the remaining classes become negligible. If the remaining probabilities are uniformly distributed, then Equations (1) and (2) yield:

$$\mathscr{P}_N = (p_{max})^N + (M-1)\left(\frac{1-p_{max}}{M-1}\right)^N, \tag{3}$$

and

$$\mathscr{P}_{N-1} = N(p_{max})^{N-1}(1-p_{max}) + N(M-1)\left(\frac{1-p_{max}}{M-1}\right)^{N-1}\left(1-\frac{1-p_{max}}{M-1}\right). \tag{4}$$

By comparing the empirical probabilities obtained for unanimous and majority agreement with Equations (3) and (4), the value of p_{max} can be estimated. A larger value of p_{max} indicates a larger level of confidence of the annotations.

Data Records

The RIVA[13] dataset is available at (https://doi.org/10.5281/zenodo.17288879) and comprises 959 image patches extracted from 115 anonymized Pap smear slides, sourced from the permanent repository of the Pathology Department at Hospital Bernardino Rivadavia, Buenos Aires, Argentina. Each image file follows the naming convention:

- `sample-category_sample-number_mini-patch-number.png`
- `sample-category`: the diagnostic category of the whole-slide sample (e.g., "HSIL", "ASCUS")
- `sample-number`: a unique identifier for the slide within that category
- `mini-patch-number`: a unique identifier for the specific patch extracted from the slide

For example, `HSIL_1_1.png` and `HSIL_1_2.png` correspond to two different image patches extracted from the same slide ("sample 1") with a global diagnosis of HSIL. Annotations are provided for each patch and include:

- the (x, y) coordinates of the nucleus center,
- the cell type classification assigned by one to four independent annotators.

Data Overview

The dataset contains a total of 26,158 annotations corresponding to 15,949 distinct cells. These annotations span eight diagnostic categories: (Pre)cancerous categories: SCC (1,586), HSIL (1,835), ASCH (416), LSIL (3,048), ASCUS (356). Non-lesion categories: INF (8,190), ENDO (1,270), NILM (9,457). Here, INF and ENDO refer to inflammatory and endocervical cells, respectively. This class distribution provides an overview of the representation of cell types within the dataset (see Fig. 1). All mini-patches and their associated metadata are available through the project's website: https://beta-digitalpapsdb.exactas.uba.ar/.

Technical Validation

To the best of our knowledge, this is the first database to enable a quantitative analysis of inter-annotator agreement-an aspect not previously addressed in the literature. The consensus study followed these steps: First, we combined annotations from all doctors to identify individual cells locations. Once the cells were identified, we examined the labels assigned to each of them. Finally, we defined a metric to quantify the degree of consensus and assess the divergence among annotations. This analysis was performed on approximately 40% of our full set of 26,158 annotations, corresponding to a subset of 400 images that were independently reviewed by all four pathologists. The remaining samples were distributed among the same four pathologists, such that each image was annotated by only one of them. As a result, these samples were excluded from the consensus analysis.

Clustering Annotations for Consensus Analysis. Before any agreement analysis could be carried out, it was essential to determine which annotations corresponded to the same underlying cell and to identify a single representative location that reflected the combined input from all annotators. This task was non-trivial for two main reasons: first, within each mini-patch, pathologists did not always annotate the exact same subset of cells; second, even when annotations referred to the same cell, their marked positions often differed slightly. To address these challenges, we employed the *MeanShift* clustering algorithm[15], which identifies groups of nearby annotations likely to refer to the same cell, thereby enabling their consolidation (see Fig. 2) Details on the clustering procedure are described in Methods. The resulting clustered dataset includes 15,949 unique cells, each annotated by one (71%), two (8%), three (8%) or four (13%) different pathologists. Figure 2 summarize the clustering statistics.

Inter-Annotator Agreement Study. To analyze inter-annotator agreement, we focused on the subset of 2,156 cells that were annotated by all four pathologists. To standardize the annotations and facilitate comparison, we defined two levels of category grouping. The first level preserves all eight original diagnostic categories. The second level consolidates these into two broader classes: pathological (SCC, HSIL, ASCH, LSIL, ASCUS) and non-pathological (INF, ENDO, NILM). This multi-tiered classification enables more flexible model tuning, allowing tolerance to annotation discrepancies based on case severity. For quantifying consensus, each annotation category was mapped to a numerical

value-1 to 8 for the full set, and 1 to 2 for the binary grouping. This numeric representation allowed us to compute dispersion metrics within each cell cluster, providing a more precise estimation of the consensus coefficient.

Based on the probabilistic model described in Methods, we estimated the expected level of consensus under random annotation conditions and compared these theoretical values to those observed in our dataset. In the binary classification case, the probability of full agreement among four annotators by chance is 12.5% (2 out of 16 possible label combinations). In contrast, our experts reached a 75% agreement rate (1627 out of 2,156 cells), a substantial increase over the random baseline. In the eight-class scenario, the expected agreement by chance is merely 0.2% (8 out of 84 possible class combinations). In contrast, the observed agreement in our dataset reached 30% (647 out of 2,156 cells)-over two orders of magnitude higher than random-underscoring the consistency and reliability of expert annotations.

To further contextualize our findings, we established a random baseline for comparison. Specifically, we simulated a dataset consisting of 2,156 annotation clusters, each assigned four categorical labels drawn uniformly at random from the eight diagnostic classes. The confusion matrices showed in Fig. 2 illustrates the extent to which expert annotations deviate from random labeling. Nearly 30% of the cells were assigned the same class by all four pathologists. Among the remaining cases, disagreements were typically confined to closely related categories, such as INF vs. ASCUS or ASCH vs. HSIL, indicating a high degree of annotation consistency even in the absence of unanimous agreement. Notably, endocervical cells, owing to their distinctive morphological characteristics, exhibited the highest rate of complete consensus among annotators.

Annotations Confidence. In this section we quantify the confidence associated with each individual annotation. To this end, we assumed that all pathologists have the same level of expertise and proposed a probabilistic model to estimate the confidence level for each annotation. This model takes into account the agreement between annotators and treats each annotation as a probabilistic outcome. Higher agreement among the annotators increases the inferred confidence for the corresponding class label, whereas greater disagreement reduces it. This approach allows us to move beyond a simple majority vote and provides a more nuanced measure of the reliability of each label.

The model, described in Methods compares empirical agreement probabilities, $\mathscr{P}$, with theoretical expectations derived from Equations (3) and (4). For the RIVA database, which includes annotations from four experts ($N = 4$), across eight diagnostic ($M = 8$), and a total of 2,156 cells annotated by all four pathologists. Among these, 647 cells exhibited unanimous agreement, and 827 showed three-to-one agreement, corresponding to empirical probabilities of $\mathscr{P}_N = 0.30$ and $\mathscr{P}_{N-1} = 0.38$ respectively. Matching these probabilities using Equations (3) and (4), implies that the maximum class probability $p_{\max}$ must be at least 0.74.

A similar analysis was conducted for the binary classification task, where the same 2,156 cells were grouped into pathological vs. non-pathological categories. In this setting, 1,627 cells received identical labels from all four annotators, while 369 showed three-to-one agreement, yielding $\mathscr{P}_N = 0.75$ and $\mathscr{P}_{N-1} = 0.17$. These values correspond to an estimated reliability of 0.94.

Together, these results demonstrate a high degree of annotation reliability in the RIVA dataset, validating both the consistency of expert input and the effectiveness of the clustering methodology. Furthermore, the estimated agreement levels offer a quantitative benchmark for expected human performance, providing valuable context for assessing future AI models trained on this task.

Usage Notes

The dataset is provided in standard image formats (PNG) together with annotation files in JSON formats, which can be easily imported into common image analysis pipelines (e.g., Python, MATLAB, ImageJ/Fiji).

Ethics Information. The project has been approved by the hospital ethics committee (IRB): Departamento de Investigación y Docencia, Hospital Bernardino Rivadavia, Buenos Aires, Argentina, by the protocol number 2023 814 GCABA HBR. The need for informed consent was waived by the IRB. The protocol established that informed consent was not required because anonymized samples from an existing repository were used. No patient-identifiable information was used in this work. The cytology slides were scanned by hospital medical staff and fully anonymized before being used for annotation and database generation.

Data availability

The RIVA[13] dataset is available at (https://doi.org/10.5281/zenodo.17288879). The dataset can also be explored in (https://beta-digitalpapsdb.exactas.uba.ar/), which enables the download of individual mini-patches and annotations, as well as the complete dataset. Final clustering results derived from the raw dataset are available at (https://github.com/LIA-DiTella/RIVA).

Code availability

The source code is publicly available at LIA-DiTella/RIVA under the GNU General Public License v3.0. It includes the scripts used to process and format the raw annotation files exported from Label Studio, which are necessary for generating the final clustering of the annotations. It also contains the script used to extract and generate the annotated mini patches. In addition to the code, the repository provides access to the raw annotation files and the final clustering results derived from them. Several additional scripts are included to facilitate the analysis and visualization of the images along with their corresponding annotations. All scripts are implemented in Python, and instructions for running the code are provided in the README section of the repository. The full-resolution scanned images in SVS format are not publicly available due to size and access restrictions; however, five representative SVS files are available to download from a link in the repository.

Received: 4 June 2025; Accepted: 5 November 2025;
Published online: 09 December 2025

References

1. Sung, H. *et al*. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **71**(3), 209–249, https://doi.org/10.3322/caac.21660 (2021).
2. Sasieni, P., Adams, J. & Cuzick, J. Benefits of cervical screening at different ages: Evidence from the UK audit of screening histories. *British Journal of Cancer* **89**(1), 88–93, https://doi.org/10.1038/sj.bjc.6601027 (2003).
3. Boone, J. D., Erickson, B. K. & Huh, W. K. New insights into cervical cancer screening. *Journal of Gynecologic Oncology* **23**(4), 282 (2012).
4. Liu, X. *et al*. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health* **1**(6), e271–e297 (2019).
5. Matias, A. V. *et al*. What is the state of the art of computer vision-assisted cytology? A systematic literature review. *Computerized Medical Imaging and Graphics* **91**, 101934 (2021).
6. Zhang, X. *et al*. A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening. *Scientific Data* **12**, 23 (2025).
7. Jantzen, J., Norup, J., Dounias, G. & Bjerregaard, B. Pap-smear benchmark data for pattern classification. *Nature Inspired Smart Information Systems (NiSIS 2005)*, 1–9 (2005).
8. 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. *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 3144–3148 (2018).
9. Rezende, M. T. *et al*. CRIC searchable image database as a public platform for conventional pap smear cytology data. *Scientific Data* **8**, 93, https://doi.org/10.1038/s41597-021-00933-8 (2021).
10. Kupas, D. *et al*. Annotated Pap cell images and smear slices for cell classification. *Scientific Data* **11**, 743 (2024).
11. Jiang, P. *et al*. A systematic review of deep learning-based cervical cytology screening: from cell identification to whole slide image analysis. *Artificial Intelligence Review* **56**(S2), 2687–2758, https://doi.org/10.1007/s10462-023-10588-z (2023).
12. Sambyal, D. & Sarwar, A. Recent developments in cervical cancer diagnosis using deep learning on whole slide images: An overview of models, techniques, challenges and future directions. *Micron* **173**, 103520, https://doi.org/10.1016/j.micron.2023.103520 (2023).
13. Perez Bianchi, P. *et al*. RIVA: An Image Dataset of Conventional Pap Smear Cytology with Multiple Independent Annotations [Data set]. Zenodo. RIVA Database https://doi.org/10.5281/zenodo.17288879 (2024).
14. Tkachenko, M., Malyuk, M., Holmanyuk, A. & Liubimov, N. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio (2020).
15. Comaniciu, D. & Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(5), 603–619 (2002).
16. Nayar, R. & Wilbur, D. C. The Bethesda system for reporting cervical cytology: A historical perspective. *Acta Cytologica* **61**(4-5), 359–372 (2017).

Acknowledgements

We thank Dr. Rubén Salanova (Biomakers) for access to scanner equipment and facilities, Diego Brunetti for assistance with database upload, and the CLIAS team for support with licensing and valuable feedback during development. We acknowledge the financial support from the Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina (grant "+4i", 2023 edition), and from CLIAS (Centro de Inteligencia Artificial y Salud para América Latina y el Caribe). CLIAS is an initiative led by the Centro de Implementación e Innovación de Políticas de Salud (CIIPS) at the Instituto de Efectividad Clínica y Sanitaria (IECS) in Argentina, with support from the International Development Research Centre (IDRC) of Canada.

Author contributions

V.S., E.I. and L.B. conceived the work. P.P.B., S.A. and M.V.C. generated the database. P.P.B. and M.A. curated the database. S.A. and M.V.C. developed the webpage. R.V.S. and J.M. prepared the samples. J.R., J.M., E.U. and L.B. scanned the samples. J.M., E.U., A.D.S. and N.B. annotated images. B.V. and H.M. developed first versions of the codes. P.P.B., A.M. and L.B. analysed inter-annotators agreement. P.P.B., A.M., V.S., E.I. and L.B. wrote the paper. E.I. made the figures. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.B.

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.