CytoCrowd: A Multi-Annotator Benchmark Dataset for Cytology Image Analysis

Anonymous Author(s)

Abstract

High-quality annotated datasets are crucial for advancing machine learning in medical image analysis. However, a critical gap exists: most datasets either offer a single, clean ground truth, which hides real-world expert disagreement, or they provide multiple annotations without a separate gold standard for objective evaluation. To bridge this gap, we introduce CytoCrowd, a new public benchmark for cytology analysis. The dataset features 446 high-resolution images, each with two key components: (1) raw, conflicting annotations from four independent pathologists, and (2) a separate, high-quality gold-standard ground truth established by a senior expert. This dual structure makes CytoCrowd a versatile resource. It serves as a benchmark for standard computer vision tasks, such as object detection and classification, using the ground truth. Simultaneously, it provides a realistic testbed for evaluating annotation aggregation algorithms that must resolve expert disagreements. We provide comprehensive baseline results for both tasks. Our experiments demonstrate the challenges presented by CytoCrowd and establish its value as a resource for developing the next generation of models for medical image analysis.

CCS Concepts

• **Applied computing** → **Bioinformatics**; • **Human-centered computing** → *Collaborative and social computing*.

Keywords

Medical Image Analysis; Annotation Aggregation; Crowdsourcing; Cytology; Benchmark Dataset

ACM Reference Format:
Anonymous Author(s). 2026. CytoCrowd: A Multi-Annotator Benchmark Dataset for Cytology Image Analysis. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3774904.3792891

1 Introduction

Machine learning models are now important tools in medical image analysis. They can assist doctors with diagnoses and expedite research. However, the success of these models depends heavily on large, high-quality annotated datasets. The need for accurate data is especially high in complex fields, such as cytology. Cytology images are difficult to annotate because they contain many overlapping cells with small but important differences. As a result, even expert pathologists often disagree on the exact boundaries, categories, or even the existence of certain objects.

Existing datasets do not fully address this challenge. General-purpose datasets like COCO [10] are not suitable for medical tasks, as they lack the specific complexities of cytology images. Many specialized medical datasets, such as BraTS [11], solve this by providing a single, unified ground truth. This approach is useful for basic model training, but it hides the real-world disagreements that occur between experts. This makes it difficult to develop models that can handle uncertainty.

Other datasets, like LIDC-IDRI [14], do provide annotations from multiple experts. This is useful for studying expert disagreement. However, they typically lack a separate gold-standard ground truth that has been verified by a senior expert. Without a final, trusted reference, it is difficult to reliably evaluate and compare different algorithms. As shown in Table 1, a benchmark is needed that offers **both** the raw expert disagreements and a separate, high-quality gold-standard ground truth.

To fill this gap, we introduce **CytoCrowd**, a new public dataset for cytology analysis. CytoCrowd is built to support two main research areas. The dataset contains 446 high-resolution images. Four independent pathologists annotated these images, creating 14,579 raw annotations that show real-world clinical disagreements. Importantly, a senior pathologist then reviewed all annotations to create a final gold-standard ground truth of 6,402 objects for reliable evaluation.

This structure makes CytoCrowd a flexible resource:

- For **Computer Vision** researchers, the gold-standard ground truth provides a clear benchmark for training and testing models for object detection, segmentation, and classification.
- For **Crowdsourcing** researchers, the raw expert annotations serve as a realistic testbed for algorithms that aim to combine multiple, conflicting annotations into a single, accurate result.

In summary, our contributions are:

(1) A new public dataset with both raw expert annotations and a separate gold-standard ground truth for cytology.
(2) A benchmark that captures real-world expert disagreements on object boundaries, categories, and existence.
(3) Baseline results for both computer vision and annotation aggregation methods to support future work.

2 Related Work

Our work is related to two main categories of datasets: standard benchmarks for medical computer vision, with a focus on cytology, and datasets containing annotations from multiple experts.

Datasets for Medical Computer Vision. Deep learning models require large, annotated datasets for training. Many high-quality medical imaging datasets have been created for this purpose. For example, the BraTS dataset [11] provides images for brain tumor segmentation, while the KiTS dataset [6] focuses on kidney tumor

WWW '26, Dubai, United Arab Emirates

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792891

Dataset	Domain	Raw Disagreements	Gold-Standard GT	Primary Use Case
COCO [10]	General Objects	No	Yes	CV Models
BraTS [11]	Brain Tumors	No	Yes (Consensus)	CV Segmentation
LIDC-IDRI [14]	Lung Nodules	Yes	No (Consensus serves as GT)	Nodule Analysis
CytoCrowd (Ours)	**Cytology**	**Yes (4 physicians)**	**Yes (Senior expert)**	**CV & Aggregation Benchmark**

Table 1: A comparison of CytoCrowd with other datasets. CytoCrowd is unique because it provides both raw expert disagreements and a separate gold-standard ground truth.

segmentation. These datasets have been essential for advancing the field.

A common characteristic of these benchmarks is that they typically provide a single, clean ground truth for each image. This ground truth is often created by having several experts create annotations, which are then merged into a single, final version. This approach is practical and provides a clear target for training standard segmentation or detection models.

However, this method of providing a single, pre-consolidated ground truth has a limitation. It hides the natural ambiguities and disagreements that are common in clinical practice. In reality, experts may disagree on the precise edges of a region or on its diagnostic classification. By removing this information, these datasets make it difficult to develop or evaluate models that can reason about uncertainty or are robust to variations in annotation style.

Multi-Annotator Datasets in Cytology. To address the issue of expert disagreement, some datasets provide raw annotations from multiple experts. The LIDC-IDRI dataset [14] is a well-known example, containing lung nodule segmentations from four radiologists. Another example is VinDr-CXR [12], which includes labels for chest X-rays from multiple annotators. These datasets are extremely valuable. They allow researchers to study the extent of inter-observer variability and provide a realistic basis for developing annotation aggregation algorithms. These algorithms aim to intelligently combine multiple, potentially conflicting labels into a single, more reliable result.

While these datasets are crucial for aggregation research, they often have a limitation when it comes to evaluation. They typically lack a separate, definitive gold-standard ground truth that stands apart from the initial annotators' opinions. For instance, the "true" label might be defined as the consensus of the annotators. This makes it challenging to perform a truly objective evaluation, as there is no independent reference to compare against.

Our work, CytoCrowd, is designed to fill the gaps left by both categories of datasets. It provides the raw, conflicting annotations necessary for aggregation research, while also offering the separate, gold-standard ground truth needed for clear and objective evaluation of computer vision models.

3 The CytoCrowd Dataset

The CytoCrowd dataset is the result of a seven-month collaborative project between the Hong Kong University of Science and Technology (Guangzhou)[1] and Guangzhou LBP Medicine Science & Technology Co.[2]. It contains 446 high-resolution cytology images, which are acquired using a whole-slide scanner at 40x magnification and stored in .svs format. This dataset is specifically created to provide a public benchmark for developing and evaluating annotation aggregation algorithms in the context of complex medical images.

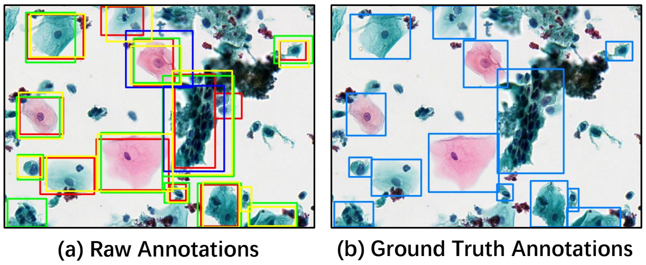
(a) Raw Annotations
(b) Ground Truth Annotations

Figure 1: Raw expert annotations (left) vs. the final gold-standard ground truth (right) on a sample image from CytoCrowd.

The annotation process is conducted by four board-certified pathologists from LBP, each possessing over ten years of clinical experience. To ensure the independence of each expert's opinion, they are asked to annotate the images separately. Using a custom-developed annotation platform[3] from the Hong Kong University of Science and Technology (Guangzhou), each pathologist identifies objects of interest by drawing bounding boxes (Regions of Interest, ROIs) and assigning a diagnostic category from a predefined set of 34 classes. This independent process generated a total of 14,579 raw annotations, capturing a wide range of inter-observer disagreements on object boundaries, categories, and existence. Figure 1 visually illustrates this core feature of our dataset. The left panel shows the raw annotations for a sample region, where different colors highlight the conflicting labels from different physicians. The right panel shows the single, consolidated gold-standard ground truth for the same region, which serves as the definitive reference for evaluation.

To create a definitive ground truth for evaluation, a senior pathologist with more than fifteen years of experience meticulously reviewed every one of the 14,579 annotations. This expert's task is to consolidate, correct, and finalize the set of ROIs and their corresponding categories for each image, establishing a gold-standard reference. This rigorous process resulted in a final ground truth of 6,402 objects.

Expert Disagreement Analysis. The complexity of the dataset is further evidenced by the inter-annotator variability. The average pairwise Intersection over Union (IoU) is 0.664, reflecting localized disagreement. Crucially, regarding object existence, only 11.37% of cells are identified by all four experts, whereas 34.78% are annotated by a single expert. This creates a high ratio of raw annotations to

[1] https://www.hkust-gz.edu.cn/
[2] https://en.gzlbp.com/
[3] https://mdi.hkust-gz.edu.cn/metal/user/

ground truth objects (approx. 2.28 to 1), underscoring the necessity of our senior-verified gold standard as a reliable benchmark. The key statistics of the CytoCrowd dataset are summarized in Table 2.

Statistic	Value
# Workers (Physicians)	4
# Tasks (Images)	446
# Total Annotations (ROIs)	14,579
# Ground Truth Objects (GT ROIs)	6,402
# Categories	34

Table 2: Statistics of the CytoCrowd Dataset.

4 Experiments and Baselines

4.1 Task Definition

The CytoCrowd benchmark is designed to evaluate methods for two distinct tasks, each targeting a different research community.

Task 1: Medical Object Detection and Classification

This is a standard computer vision task. The goal is to train a model that can directly identify the location and diagnostic category of cellular objects in the images.

- **Input:** The high-resolution cytology images.
- **Training Data:** Models are trained using the **gold-standard ground truth** annotations.
- **Goal:** For a given test image, the model should produce a set of bounding boxes and a corresponding class label for each box.
- **Evaluation:** The model's predictions are compared against the gold-standard ground truth of the test set.

Task 2: Annotation Aggregation

This task is for the crowdsourcing and truth inference community. The goal is to combine the conflicting annotations from multiple experts into a single, high-quality result that is as close to the ground truth as possible.

- **Input:** The raw, and often conflicting, annotations from the four independent pathologists for a given image.
- **Goal:** The algorithm should process multiple inputs and produce a single, consolidated set of annotations (bounding boxes and class labels).
- **Evaluation:** The algorithm's consolidated output is compared against the **gold-standard ground truth**.

By defining these two separate tasks, we provide a clear benchmark for both learning-based vision models and aggregation algorithms.

4.2 Evaluation Metrics

To evaluate performance, we report the classification **Accuracy**. This metric is calculated in a way that separates localization success from classification correctness. Specifically, we first match predicted objects to ground truth objects using the standard **IoU** metric. A prediction is considered correctly located—a **true positive (TP)**—if its IoU with a ground truth object is greater than 0.5. The final Accuracy is then calculated as the percentage of correctly classified objects *only among these TPs*. This approach ensures that the accuracy score purely reflects the model's ability to classify objects that it has successfully located, which provides a clear and fair comparison across different methods.

Method	Accuracy
CATD	0.857
Dawid & Skene (D&S)	0.893
Majority Voting (MV)	0.903
PM	0.855
LFC	0.896
ZenCrowd	0.883

Table 3: Performance of annotation aggregation methods. Accuracy is calculated on correctly localized objects (TPs).

4.3 Baseline Methods

We provide baseline results for the two tasks defined for the CytoCrowd dataset. These baselines are chosen to represent common or powerful approaches in their respective fields.

4.3.1 Annotation Aggregation Baselines. For the annotation aggregation task, we evaluate well-known inference methods including **Majority Voting (MV)**, the statistical **Dawid & Skene (D&S) [3]** model, and truth discovery methods such as **CATD [8]**, **PM [9]**, **LFC [13]**, and **ZenCrowd [4]**.

4.3.2 Learning-based Baselines. For the medical object detection and classification task, we test several modern, powerful vision models instead of traditional object detectors.

- **DeepEdit [5]** and **Anytime [7]:** These are models based on or related to interactive and prompt-based segmentation, representing the state-of-the-art in producing precise object masks.
- **Qwen-VL-MAX [1]** and **Qwen2.5-VL-72B [2]:** These are large-scale Vision-Language Models (VLMs). We test them to see if their extensive general-world knowledge can be applied to this specialized medical task.

4.4 Performance Analysis

We present the performance of the baseline methods on the two defined tasks.

4.4.1 Performance of Annotation Aggregation Methods. Table 3 shows the accuracy results for the annotation aggregation baselines. The performance is measured by comparing the aggregated category labels against the gold-standard ground truth for all correctly located objects.

The results are very informative. Interestingly, the simplest baseline, **Majority Voting (MV), achieves the highest accuracy (0.903)**, outperforming more complex models like Dawid & Skene (D&S). This finding is significant as it suggests that when all annotators are domain experts, their collective agreement provides a powerful signal. In this scenario, complex models that try to learn and correct for annotator errors may not provide an advantage, as the experts are all highly reliable. This result highlights the unique nature of our expert-annotated dataset and presents a challenge to existing aggregation methods.

4.4.2 Performance of Learning-based Methods. Table 4 presents the accuracy of the vision models on our dataset. The results show a clear and consistent trend.

There is a major performance gap between the general-purpose VLMs and the more specialized segmentation models. The **Qwen VLMs perform very poorly**, with accuracy below 45%. This

Method	Accuracy
Qwen-VL-MAX	0.441
Qwen2.5-VL-72B	0.437
DeepEdit	0.899
Anytime	0.878

Table 4: Comparison on CV/VLM baselines.

strongly indicates that despite their vast knowledge, these large models cannot effectively handle the fine-grained and domain-specific challenges of cytology image analysis without specialized fine-tuning.

In contrast, the models more focused on segmentation, **DeepEdit and Anytime, achieve high accuracy (0.899 and 0.878, respectively)**. Their strong performance establishes a solid and competitive baseline for future computer vision research. This demonstrates the CytoCrowd dataset's value as a benchmark for developing and testing new, specialized models that address challenges missed by general models.

5 Conclusion

We have introduced **CytoCrowd**, a new expert-annotated benchmark for complex medical image annotation aggregation. Our extensive benchmarking demonstrates the difficulty of the task. We acknowledge two limitations: the dataset size (446 images) is relatively small compared to general CV benchmarks, and the single-expert gold standard may introduce observer bias. Future iterations could incorporate biopsy-verified labels. Despite this, we hope CytoCrowd will spur the development of novel algorithms that can effectively harness the collective intelligence from multiple, conflicting expert annotations.

References

[1] Jinze Bai, Shuai Bai, et al. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR* abs/2308.12966 (2023).

[2] Shuai Bai, Keqin Chen, Xuejing Liu, et al. 2025. Qwen2.5-VL Technical Report. *CoRR* abs/2502.13923 (2025).

[3] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 1 (1979), 20–28.

[4] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *WWW*. ACM, 469–478.

[5] Andres Diaz-Pinto, Pritesh Mehta, et al. 2023. DeepEdit: Deep Editable Learning for Interactive Segmentation of 3D Medical Images. *CoRR* abs/2305.10655 (2023).

[6] Nicholas Heller, Niranjan Sathianathen, et al. 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445* (2019).

[7] Pranav Kulkarni, Adway U. Kanhere, Dharmam Savani, Andrew Chan, Devina Chatterjee, Paul H. Yi, and Vishwa S. Parekh. 2024. Anytime, Anywhere, Anyone: Investigating the Feasibility of Segment Anything Model for Crowd-Sourcing Medical Image Annotations. *CoRR* abs/2403.15218 (2024).

[8] Qi Li, Yaliang Li, et al. 2014. A Confidence-Aware Approach for Truth Discovery on Long-Tail Data. *Proc. VLDB Endow.* 8, 4 (2014), 425–436.

[9] Qi Li, Yaliang Li, et al. 2014. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *SIGMOD Conference*. ACM, 1187–1198.

[10] Tsung-Yi Lin, Michael Maire, et al. 2014. Microsoft COCO: Common Objects in Context. In *ECCV (5) (Lecture Notes in Computer Science, Vol. 8693)*. Springer, 740–755.

[11] Bjoern H Menze, Andras Jakab, et al. 2014. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging* 34, 10 (2014), 1993–2024.

[12] Ha Q Nguyen, Khanh Lam, et al. 2022. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data* 9, 1 (2022), 429.

[13] Vikas C. Raykar, Shipeng Yu, et al. 2010. Learning From Crowds. *J. Mach. Learn. Res.* 11 (2010), 1297–1322.

[14] G Samuel. 2011. The Lung Image Database Consortium (LIDC) and Image Database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical physics* 38 (2011), 2.