WBCAtt Datasheet

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

This dataset was developed to facilitate research on explainable and interpretable machine learning models for the classification of White Blood Cells (WBCs) based on their morphological characteristics. While there are existing datasets for WBC classification, none of them include a comprehensive set of morphological characteristics for WBCs. By creating a densely annotated dataset, we aim to fill this gap and provide researchers with a resource that can be used to develop more accurate and interpretable models for WBC recognition.

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

This dataset was built upon an existing peripheral blood cells dataset (PBC dataset) [1], which only included the cell types and was used primarily for benchmarking WBC classification algorithms. Our new dataset, WBCAtt, serves as an extension to the previous dataset, providing more comprehensive morphological annotations for WBCs, which can be used to improve the accuracy and interpretability of WBC classification models.

C. What (other) tasks could the dataset be used for?

This dataset could be used for diverse tasks beyond explainable WBC recognition. The examples include even training image generation models to synthesize WBC images with specific morphological features.

D. Who funded the creation dataset?

The creation of this annotation dataset was supported in part by Sysmex Corporation.

E. Any other comment?

None.

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The instances in this dataset are images of WBCs from the PBC dataset [1]. The original images contain information about the class of the WBC. Our new dataset is densely annotated with 11 morphological attributes established through consultation with a pathologist and a literature review.

B. How many instances are there in total (of each type, if appropriate)?

The list of images for each WBC types are tabulated in Table I. We annotated 11 attributes for these 10,298 images, resulting in 113,278 image-attribute pairs. The complete list of attributes and their values are tabulated in Table II.

TABLE I
INSTANCES OF EACH WBC TYPE.

Cell Type	Number of Images
Neutrophils	3329
Eosinophils	3117
Basophils	1218
Lymphocytes	1214
Monocytes	1420
Total	10298

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Each instance in the dataset consists of a single peripheral blood cell image, with the blood cell class labeled by clinical pathologists (provided in the original dataset) [1]. Each image in the dataset is newly annotated by the 11 attributes that describe the morphological appearance of the WBCs, which are commonly used by pathologists to determine the WBC types. There is no age or gender information associated with this dataset. The distribution of values per attribute is tabulated in Table II.

D. Is there a label or target associated with each instance? If so, please provide a description.

Yes, the images were initially labeled with the types of WBCs (neutrophil, eosinophil, basophil, lymphocyte, and

TABLE II

ATTRIBUTE DIST. THE DISTRIBUTION REPRESENTS THE RESULTS OF ANNOTATING ALL TYPICAL WBCS FROM THE PBC DATASET, WHICH IS THE IMAGE SOURCE WE UTILIZED. WE DID NOT ACTIVELY CONTROL OR MANIPULATE THE DISTRIBUTION.

Attribute	Value (Count)
Cell-Size	Big (4,997), Small (4,271)
Cell-Shape	Round (7,173), Irregular (2,095)
Nucleus-Shape	Segmented-Bilobed (2,806), Unsegmented-Band (2,356), Unsegmented-Indented (1,205), Segmented-Multilobed (1,143), Unsegmented-Round (967), Irregular (791)
Nuclear-Cytoplasmic-Ratio	Low (8,148), High (1,120)
Chromatin-Density	Densely (8,443), Loosely (825)
Cytoplasm-Vacuole	No (8,559), Yes (709)
Cytoplasm-Texture	Clear (7,429), Frosted (1,839)
Cytoplasm-Color	Light Blue (7,011), Blue (1,273), Purple Blue (984)
Granularity	Yes (6,896), No (2,372)
Granule-Type	Small (3,003), Round (2,801), Coarse (1,090), Nil (2,374)
Granule-Color	Pink (2,925), Red (2,803), Purple (1,167), Nil (2,373)

monocyte) [1]. Later, we added 11 morphological attribute labels to each image, which were developed through the discussions with pathologists, review of the relevant literature, and manual inspection of WBC images. Table II shows the complete list of attributes and their values. The detailed description of each attributes are provided in the main paper, Section 3.

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The dataset does not include information on the number of blood smear collected, as well as the age and gender of the patients from whom the samples were taken. This information was not available from the original dataset.

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Not applicable.

G. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

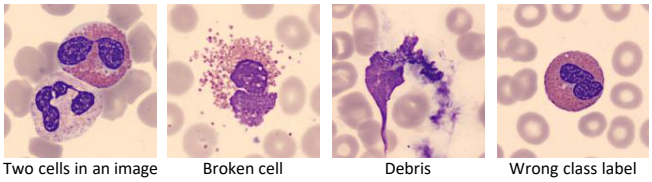

Fig. 1. Potential noises exist in the attribute annotations. The cell type label for the rightmost image is incorrectly assigned as *neutrophil*, whereas it should be *eosinophil*.

The dataset represents a sample rather than containing all possible instances from a larger set. Specifically, the dataset used for this study includes morphological attribute annotations for the five major types of white blood cells (neutrophils, eosinophils, basophils, lymphocytes, and monocytes) collected in a particular hospital. Defining the larger set precisely is challenging as it comprises cell images from diverse locations worldwide, representing various patient conditions, races, and species of mammals, while also incorporating a wide range of employed staining methods. The dataset [1] we used is from the Hospital Clinic of Barcelona, located in Barcelona, Catalonia, Spain. They collected images of normal peripheral blood cells, which were obtained from samples collected in their Core Laboratory. May Grünwald-Giemsa staining was used to stain the cells. The dataset comprises images from normal human individuals, and the selection of blood cells was based on normal laboratory data. These images were collected over a 4-year period, from 2015 to 2019, as part of the hospital's daily routine.

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The dataset is randomly split into 6,179 training images, 1,030 validation images, and 3,099 test images. The random division ensures that the cell-type distributions are the same in each set. This allows for reliable evaluation and comparison of different models including those from future work.

I. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Our dataset is built upon the existing WBC image dataset [1]. During the dataset construction process, we encountered a limited number of instances where incorrect cell type labels were provided by the PBC dataset. Additionally, we observed the presence of images containing two WBCs within a single image, as well as cells that appeared to be broken. These factors have the potential to introduce errors or inconsistencies in the attribute annotations. Figure 1 shows these examples.

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Our dataset is developed from the existing PBC dataset [1]. The PBC dataset was developed and is maintained by Andrea Acevedo and colleagues in Hospital Clinic de Barcelona. Their work is licensed under a Creative Commons Attribution 4.0 International license. The PBC dataset can be accessed here: `https://data.mendeley.com/datasets/snkd93bnjr/1`. There is no cost to use the dataset for non-commercial research and educational purposes.

Any other comments? None.

III. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The processes used to collect the WBC images and their cell type annotations are described in [1]. The attribute annotation process involved a combination of manual human curation and expert input from pathologists. Initially, discussions with pathologists from a healthcare company provided the basis for identifying five prominent attributes related to the major WBC types. These attributes were refined through literature review and further discussions with pathologists. Approximately a thousand WBC images were manually inspected to finalize the set of 11 attributes, each supported by medical literature references.

To ensure reliable annotations, a rigorous annotation process was implemented. Biomedical students annotated the images, being informed of the specific WBC type for each image. A screenshort of the annotation interface is shown in Figure 2. The annotations were then meticulously reviewed by research scientists, which means that each image is examined by at least two individuals. Any ambiguities were resolved through discussions with the pathologists who defined the attributes with us. Quality control measures were implemented, and the details are available in the Appendix.

For further validation, a subset of 1,000 images was annotated independently by different annotators. The results showed a high agreement rate, with 10,569 out of 11,000 attribute annotations being consistent with the original annotations, resulting in an agreement rate of approximately 96.1%.

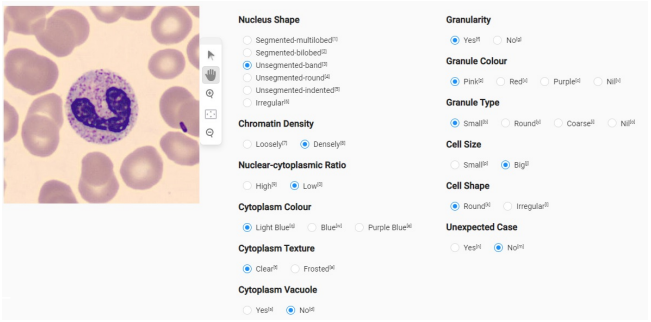

Fig. 2. Screenshot of the annotation interface.

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The WBC images were labeled by biomedical students, and the labels were subsequently reviewed and assessed by research scientists.

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

All images from the five major types of WBCs in the PBC dataset (neutrophils, eosinophils, basophils, lymphocytes and monocytes) were used for this work.

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The biomedical students from Nanyang Technological University (in Singapore) were involved in labeling of the data. These students were compensated based on the number of images they labeled, with an average payment of 0.17 SGD per image. On average, the students labeled 180 images per hour, which translates to an hourly compensation of about 30 SGD, which is more than three times the minimum wage in Singapore [2].

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

According to the original paper [1], the collection of WBC images took place over a period spanning from 2015 to 2019. The attribute annotations were conducted in 2023.

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No preprocessing. The labeling of the dataset was done manually by human annotators.

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Not applicable.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Not applicable.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Not applicable.

E. Any other comments

None.

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The PBC dataset can be accessed through the following link: `https://data.mendeley.com/datasets/snkd93bnjr/1`, and has been assigned a DOI: 10.17632/snkd93bnjr.1. Additionally, the newly annotated morphological attribute dataset, WBCAtt, is uploaded here: `https://github.com/apple2373/wbcatt`. There is no redundant archive for this dataset at the current moment, but upon acceptance of the paper, we will upload to the `data.mendeley.com`, which gives an DOI.

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset is already available. The PBC dataset is licensed under a Creative Commons Attribution 4.0 International license. The WBCAtt annotations are licensed under MIT.

C. Are there any copyrights on the data?

See above for licensing of datasets.

D. Are there any fees or access/export restrictions?

There are no fees. The PBC dataset is intended to be used for research and educational purposes only. Our dataset is also intended to be used for research and educational purposes.

E. Any other comments?

None.

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The dataset and the website where the annotations are released will be maintained by the authors of the manuscript.

B. Will the dataset be updated? If so, how often and by whom?

In the future, there may be updates and expansions to the dataset but we do not have any concrete update planed yet. Any updates or changes to the dataset will be communicated and posted on the dataset webpage, if available.

C. How will updates be communicated? (e.g., mailing list, GitHub)

Updates will be communicated through the dataset website: `https://github.com/apple2373/wbcatt`

D. If the dataset becomes obsolete how will this be communicated?

Through the dataset website: `https://github.com/apple2373/wbcatt`

E. Is there a repository to link to any/all papers/systems that use this dataset?

There is a repository, maintained by the authors of the manuscript at `https://github.com/apple2373/wbcatt`.

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

The dataset is under the MIT licence so anyone has freedom to do so. Currently, we do not have mechanisms in place; however, others may contact us to discuss potential use cases if they prefer.

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Our study was exempt from institutional IRB approval as we added labels to publicly available data.

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

The WBC images in PBC dataset [1] were saved with a random identification number to remove any link and traceability to the patient data, resulting in an anonymized dataset. The morphorlogical attributes in WBCAtt describe what is seen in these WBC images.

C. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

The dataset contains cell images which may potentially cause distress to some individuals.

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes, the dataset contains WBC images from human peripheral blood smears.

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

There is no age or gender information contained in the dataset.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

The dataset only includes deidentified images. The WBC images in PBC dataset [1] were saved with a random identification number to remove any link and traceability to the patient data, resulting in an anonymized dataset.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The WBCAtt dataset utilizes existing public dataset.

I. Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The consent process for the PBC dataset [1] was not specifically described by the authors. However, the dataset is published in a journal with rigorous ethical guidelines[1], and their study is approved by their institutional IRB, indicating that it was conducted in an ethical manner.

J. Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

See above.

K. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Not applicable.

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Such an analysis has not been completed.

M. Any other comments?

None.

VIII. Aknowledegment

The template for this datasheet was obtained from `https://www.overleaf.com/latex/templates/datasheet-for-dataset-template/ztkyvzddvxtd` and slightly modified for our purposes.

References

[1] Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30, 2020.

[1] `https://www.sciencedirect.com/journal/data-in-brief/about/policies-and-guidelines`

[2] Ministry of Manpower Singapore. Local Qualifying Salary. https://www.mom.gov.sg/employment-practices/progressive-wage-model/local-qualifying-salary, 2022. Accessed: 2023-06-12.