Motivation

For what purpose was the dataset created?

With Raw-Microscopy we provide a publicly available raw image dataset in order to examine the effect of the image signal processing on the performance and the robustness of machine learning models. This dataset enables to study these effects for a supervised multiclass classification task: the classification of white blood cells (WBCs).

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

This dataset has been created by the Laboratory of Applied Optics of the Micro-Nanotechnology group at HEPIA/HES-SO, University of Applied Sciences of Western Switzerland. Single-cell images were annotated by a trained cytologist.

Who funded the creation of the dataset?

The creation of the dataset has been funded by HEPIA/HES-SO.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

An instance is a tuple of an image and a label. The image shows a human WBCs and the label indicates the morphological class of this cell. The following eight morphological classes appear in the dataset: Basophil (BAS), Eosinophil (EOS), Smudge cell (KSC), atypical Lymphocyte (LYA), typical Lymphocyte (LYT), Monocyte (MON), Neutrophil (NGB), Neutrophil (NGS). The nith class consists of images that could not be assigned a class (UNC) during the labeling process.

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

The data set consists of 940 instances. For the proportion of each class in the dataset see table 6.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset does not contain all possible instances. It is limited to WBC classes normally present in the peripheral blood of healthy humans. In order to cope with intrinsic class imbalance in cell distribution, rare cell class candidates such as Basophils were perferentially imaged.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Each instance consists of an image of 256 by 256 pixels. The image is a raw image in .tiff format.

Is there a label or target associated with each instance?

Each instance is associated to a label, that indicates the morphological class of the image.

Is any information missing from individual instances?

No information is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

No, relationships between indivduals are not made explicit.

Are there recommended data splits (e.g., training, development/validation, testing)?

There are no recommended data splits. All the data splits that we used for our experiments were randomly picked.

Are there any errors, sources of noise, or redundancies in the dataset?

To the best of our knowledge, there are no errors in the dataset. However, a key source of variability between slides from different laboratories and processing times is stain intensity. The samples used in this work all come from the same source, hence we assume the preanalytic treatment and staining protocol to be similar. As all images were obtained on the same microscopy equipment, focus handling and illumination are identical for all samples. Image labelling was performed by one trained morphologist with experience in hematological routine diagnostics. It is known that morphology annotations are subject to inter- and intra-rater variability. However, as we limit ourselves to normal WBCs the labeling is expected to be stable.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?

The dataset consist of medical data, disclosing the morphological classes of single human WBCs. In principle, the distribution of cell types conveys information on the health state of a patient. However, the subjects in this dataset are fully de-identified, so that the image data cannot be linked back to the healthy donors of the scanned blood smears. Furthermore, it is not disclosed which cell image was taken from which blood smear, so that no frequencies of individual cell types can be determined. Additionally, we only consider cell types present in normal blood, so that no specific hematologic pathology can be deduced from cell morphologies.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No. The dataset does not contain data with any of the above properties.

Does the dataset relate to people?

Yes. The dataset consist of images of human WBCs.

Does the dataset identify any subpopulations (e.g., by age, gender)?

The donors of the blood smears used in this dataset are fully deidentified, and no information on subpipulation composition is provided.

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?

No. It is not possible to identify individuals from an image of their white blood cells or visa versa.

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)?

No. While the distribution of cell types for a specific patient could reveal information about that patient's health status, isolated single-cell images of normal leukocytes do not allow for this inference.

Any other comments?

See table 7 for a summary of the composition of Raw-Microscopy.

class	proportion in %
Basophil (BAS)	1.91
Eosinophil (EOS)	5.74
Smudge cell / debris (KSC)	17.34
Lymphocyte (LYA)	3.19
Lymphocyte (LYT)	24.47
Monocyte (MON)	20.32
Neutrophil (NGB)	0.85
Neutrophil (NGS)	22.98
Image that could not be assigned a class (UNC)	3.19

Table 6: The proportion of the classes in RawSet-Microscopy.

Collection Process

How was the data associated with each instance acquired?

Images of the dataset have been acquired directly from a CMOS imaging sensor. They are in a raw unprocessed format.

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

Imaging data have been obtained via a custom brightfield microscope.

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

Images have 256×256 pixel size and have been cropped from larger images. The dataset corresponds to a selection of white blood cells in the acquired large images. A sampling strategy aimed at increasing the proportion of rare classes of white blood cells has been used.

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)?

A research assistant has been involved in the data collection process and has been compensated with a monthly salary.

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)?

Data have been collected on a timeframe of two months, corresponding to the availability of the physical samples to image. Data have been collected on purpose for this work.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No, imaging data have been obtained from blood smear slides slides bought from a private company (J. Lieder GmbH Co. KG, Ludwigsburg/Germany).

Does the dataset relate to people?

Yes. The dataset consists of microscopic images of human white blood cells.

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

Data have not been obtained via third parties.

Were the individuals in question notified about the data collection?

As the blood smear slides were bought from a company, notification to individuals of the data collection has been performed by the company.

Did the individuals in question consent to the collection and use of their data?

Yes, they did.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

We do not know the conditions of consent adopted by the selling company. However, we believe the company provided the individuals a complete freedom in revoking their consent in the future, if desired.

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?

No, this kind of analysis has not been conducted.

Preprocessing/cleaning/labeling

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)?

Intensity scaled images are generated with Jetraw Data Suite for both datasets, which applies a physical model based on sensor calibration to accurately simulate intensity reduction. Microscopy Raw images are extracted from RGB Microscopy data through a pixel selection from images taken with three filters, in order to have a Bayer Pattern. Pixels intensities are rescaled with Jetraw Data Suite to match the measured transmissivities of a Bayer coulour filters aray.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Raw images are available in the dataset.

Is the software used to preprocess/clean/label the instances available?

At the time of writing, the software is not available, it will be made available in the months to come.

Uses

Has the dataset been used for any tasks already?

The dataset has not yet been used.

Is there a repository that links to any or all papers or systems that use the dataset?

The repository at `https://aiaudit.org/lens2logit/` associated to this work, maintained by Luis Oala.

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

The dataset can be used to study the effect of image signal processing on the performance and robustness of any other machine learing model implemented in PyTorch, designed for a supervised multiclass classification task.

Is there anything about the composition of the dataset or the way it was col-

lected and preprocessed/cleaned/labeled that might impact future uses?

To the best of our knowledge, we do not recognize such impacts.

Are there tasks for which the dataset should not be used?

To the best of our knowledge, there are no such tasks.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes. The dataset will be publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)

A guide to access the dataset is available at `https://aiaudit.org/lens2logit/`. Moreover, the dataset can be downloaded directly at `https://zenodo.org/` under the doi: 10.5281/zenodo.5235536.

When will the dataset be distributed?

The dataset is already publicly available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset will be distributed under the Creative Commons Attribution 4.0 International.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

There are no such restrictions.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Bruno Sanguinetti on behalf of Dotphoton AG.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

By email address via
bruno.sanguinetti@dotphoton.com.

Is there an erratum?

At the time of submisson, there is no such erratum. If an erratum is needed in the future it will be accessible at `https://aiaudit.org/lens2logit/`.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes. The dataset will be enlarged wrt. the number of instances.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

To the best of our knowledge, there are no such limits.

Will older versions of the dataset continue to be supported/hosted/maintained?

Older versions will be supported and maintained in the future. The dataset will continue to be hosted as long as `https://zenodo.org/` exists.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

For any of these requests contact either Luis Oala (luis.oala@hhi.frauhofer.de) or Bruno Sanguinetti (bruno.sanguinetti@dotphoton.com). For now, we do not have an established mechanism to handle these requests.

Composition of Raw-Microscopy	
Type of instances	image and label
Objects on images	white blood cells
Type of classes	morphological classes
Number of instances	940
Number of classes	9
Image size	256 by 256 pixels
Image format	tiff

Table 7: A summary of the composition of Raw-Microscopy.

Motivation

For what purpose was the dataset created?

With Raw-Drone we provide a publicly available raw dataset in order to examine the effect of the image data processing on the performance and the robustness of machine learning models. This dataset enables to study these effects for a segmentation task: the segmentation of cars. The dataset was taken with specified parameters: sensor gain, point-spread function and ground-sampling distance, so that physical models may be used to process the data. It also was taken with a easily accessible and affordable system, so that it may be reproduced.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Bruno Sanguinetti on behalf of the company Dotphoton AG.

Who funded the creation of the dataset?

The data collection was funded by Dotphoton AG. The calibration of the imager characteristics was jointly funded by Dotphoton AG and the European Space Agency.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

An instance is a tuple of an image and a segmentation mask. The image shows a landscape shot from above. The segmentation mask is a binary image. A white pixel in this mask corresponds to a pixel within a region in the image where a car is displayed. A black pixel in this mask corresponds to a pixel within a region in the image where no car is displayed.

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

The dataset consists of 548 instances.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The dataset does not contain all possible instances. Only images with at least one white pixel in the associated segmentation mask are considered.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Both, the image and the segmentation mask consist of 256 by 256 pixels. The image is a raw image in .tif format and the the segmentation mask is in .png format. The images are cropped sub-images of 12 raw images in .DNG format, consisting of 3648 by 5472 pixels.

Is there a label or target associated with each instance?

Each instance is associated to a binary segmentation mask.

Is any information missing from individual instances?

No information is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?

Since every image is a cropped sub-image of an original image, several of these sub-images belong to the same original image. All sub-images are disjoint, i.e. no different images share a pixel from the original image.

Are there recommended data splits (e.g., training, development/validation, testing)?

There are no recommended data splits. All the data splits that we used for our experiments were randomly picked.

Are there any errors, sources of noise, or redundancies in the dataset?

To the best of our knowledge, there are no errors in the dataset. The segmentation mask is created by hand and hence noisy, especially at the boundaries between a region with a car and a region without a car.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?

No. The dataset does not contain data of any of the above types.

Does the data set contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No. The dataset does not contain data with any of the above properties.

Does the dataset relate to people?

The dataset does not relate to people. There are individuals on the images, but it is not possible to identify these individuals.

Any other comments?

See table 8 for a summary of the composition of the Raw-Drone.

Collection Process

How was the data associated with each instance acquired?

The data was collected by flying a drone and saving the raw data. The calibration data for the drone's imager was acquired both under laboratory conditions and using a ground-based calibration target, so that it could be acquired under operating conditions.

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

To acquire the drone images, we used a DJI Mavic 2 Pro Drone, equipped with a Hasselblad L1D-20c camera (Sony IMX183 sensor). This system has $2.4\,\mu\mathrm{m}$ pixels in Bayer filter array. Images were taken with the drone hovering for maximum stability. This stability was verified to be better than a single pixel by calculating the correlation of subsequent images. The objective has a focal length of $10.3\,\mathrm{mm}$. We operated this objective at an f-number of $N = 8$, to emulate the PSF circle diameter relative to the pixel pitch and ground sampling distance (GSD) as would be found on images from high-resolution satellites. Operating at $N = 8$ also minimises vignetting, aberrations, and increases depth of focus. The point-spread function (PSF) was measured to have a circle diameter of $12.5\,\mu\mathrm{m}$ using the edge-spread function technique and a ground calibration target.This corresponds to $\sigma = 2.52\,\mathrm{px}$, which also corresponds to a diffraction-limited system, within the uncertainty dictated by the wavelength spread of the image. Images were taken at $200\,\mathrm{ISO}$, corresponding to a gain of $0.528\,\mathrm{DN}/e^-$. The 12-bit pixel values are however left-justified to 16-bits, so that the gain on the 16-bit numbers is $8.448\,\mathrm{DN}/e^-$. The images were taken at a height of $250\,\mathrm{m}$, so that the GSD is $6\,\mathrm{cm}$. All images were tiled in 256x256 patches. Segmentation color masks were created to identify cars for each patch. From this mask, classification labels were generated to detect if there is a car in the image. The dataset is constituted by 548 images for the segmentation task, and 930 for classification. The dataset is augmented through , with 7 different intensity scales.

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

The entire dataset is presented.

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 dataset was taken by a company employee, compensated by his salary. Labeling was performed by both a company employee and a PhD student, who's PhD is funded by the company.

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)?

The dataset was taken as the initial step of writing this article.

Were any ethical review processes conducted (e.g., by an institutional review board)?

The dataset does not contain any elements requiring an ethical review process.

Does the dataset relate to people?

The dataset does not relate to people. There are individuals on the images, but it is not possible to identify these individuals.

Preprocessing/cleaning/labeling

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)?

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Is the software used to preprocess/clean/label the instances available?

Uses

Has the dataset been used for any tasks already?

The dataset has not yet been used.

Is there a repository that links to any or all papers or systems that use the dataset?

The repository at `https://aiaudit.org/lens2logit/` associated to this work, maintained by Luis Oala.

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

The dataset can be used to study the effect of image signal processing on the performance and robustness of any other machine learing model implemented in PyTorch, designed segmentation task.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

To the best of our knowledge, we do not recognize such impacts.

Are there tasks for which the dataset should not be used?

To the best of our knowledge, there are no such tasks.

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?

Yes. The dataset will be publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)

A guide to access the dataset is available at `https://aiaudit.org/lens2logit/`. Moreover, the dataset can be downloaded directly at `https://zenodo.org/` under the doi: 10.5281/zenodo.5235536.

When will the dataset be distributed?

The dataset is already publicly available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The dataset will be distributed under the Creative Commons Attribution 4.0 International.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

There are no such restrictions.

Maintenance

Who will be supporting/hosting/maintaining the dataset?

Bruno Sanguinetti on behalf of Dotphoton AG.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

By email address via
bruno.sanguinetti@dotphoton.com.

Is there an erratum?

At the time of submisson, there is no such erratum. If an erratum is needed in the future it will be accessible at `https://aiaudit.org/lens2logit/`.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?

Yes. The dataset will be enlarged wrt. the number of instances.

Will older versions of the dataset continue to be supported/hosted/maintained?

Older versions will be supported and maintained in the future. The dataset will continue to be hosted as long as `https://zenodo.org/` exists.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?

To the best of our knowledge, there are no such limits.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?

For any of these requests contact either Luis Oala (luis.oala@hhi.frauhofer.de) or Bruno Sanguinetti (bruno.sanguinetti@dotphoton.com). For now, we do not have an established mechanism to handle these requests.

Composition of Raw-Drone	
Type of instances	image and mask
Objects on images	landscape shots from above
Number of instances	548
Number of original images	12
Image size	256 by 256 pixels
Mask size	256 by 256 pixels
Original image size	3648 by 5472
Image format	tif
Mask format	png
Original image format	DNG

Table 8: A summary of the composition of Raw-Drone.