DermX supplementary material

This document incorporates all supplementary material associated with the DermX dataset. DermX annotations can be found at https://github.com/ralucaj/dermx, along with instructions on how to read it and how to access the original images from DermNetNZ and SD-260.

Hosting, licensing, and maintenance plan

The dataset is currently stored on GitHub. Once the paper is accepted, it will be moved to DTU Data to facilitate its hosting and maintenance.

DermX annotations are available under the Creative Commons licence. The original images are the property of DermNetNZ and SD-260, and inherit their respective dataset's licence. The owners of the two datasets must be contacted to gain permission to use the images.

DermX will be routinely updated, either when needed if errors are found and corrected, or by creating a new release. The latter option will happen on the addition of dermatologist evaluations, target conditions, or on updates to the ontology.

Author responsibility

The main author bears all responsibility in case of violation of rights. The dataset annotations are available under the Creative Commons licence, while the original images remain the property of DermNetNZ and SD-260, respectively. Original images are also available under the Creative Commons licence, but approval from the authors must be obtained before using them.

Datasheets for Datasets

In this supplementary material, we follow the Datasheets for Datasets framework to future-proof the usage of the DermX dataset.

Motivation

For what purpose was the dataset created?

The dataset was created to provide a gold standard annotation for skin lesion diagnosis explainability. This enables researchers to objectively compare ConvNets in terms of explainability, by comparing the ConvNet explanations to the explanations provided by expert dermatologists.

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

The dataset was created by the Machine Learning group at Omhu, in collaboration with the Cognitive Systems at the Technical University of Denmark (DTU).

Who funded the creation of the dataset?

The dataset creation was funded by Omhu and the Innovation Fund Denmark.

Composition

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

The dataset consists of JSON evaluations files performed by expert dermatologists for publicly available images stored, and PNG segmentations of characteristics that support the dermatologists' diagnosis.

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

The dataset consists of 1285 evaluations and 3562 segmentations.

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

Images were randomly sampled from two publicly available datasets, DermNetNZ and SD-260 such that they represent the six target conditions.

What data does each instance consist of?

Evaluations are stored as JSON files, and contain information about the image they are associated with, and diagnosis and characteristics tagged on the image. If characteristics were localisable, the files also contain the bounding box and the complex polygon paths to recreate the masks. If additional descriptive terms could be added, they are also listed. Characteristic masks are stored as PNG masks.

Is there a label or target associated with each instance?

All labels are available in the csv files that further explain the dataset.

Is any information missing from individual instances?

No.

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

Yes, in the associated instance_mask_annotations.csv and metadata.csv files.

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

No.

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

Original images were associated with their classes based on the author's understanding of the disease names (e.g. palmo-plantar psoriasis was tagged as psoriasis). With regards to the annotations, the dataset is subject to human annotator error and noise.

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

The dataset links to two public dataset, DermNetNZ and SD-260. There are no guarantees that these datasets will not change in the future, outside of the owners interest to keep maintaining them. No official archival versions of the complete dataset are available at the moment, as the author did not yet receive the rights to redistribute the original images. Both external datasets are available under the Creative Commons licence.

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 data introduced by this dataset has no confidential information associated with it.

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

The added data has no sensitive data. The original images contain pictures of skin diseases that may cause anxiety.

Does the dataset relate to people?

Yes.

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

The dataset includes annotations of age, biological sex, and skin tone in broad terms (e.g. young, dark skin), when deemed relevant as an explanation for the diagnosis. Not all images have this annotation, as sometimes it was not considered a good explanation. Proper distributions of these subpopulations cannot be obtained.

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?

Within the additional annotations, no. Within the original images, several photos display patient faces.

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)?

Within the additional annotations, no. Within the original images, the photos that display faces may reveal racial or ethnic origins.

Collection process

How was the data associated with each instance acquired?

Dermatologists were instructed to annotate images using the annotation protocol described in the paper. Each image was annotated by three dermatologists in Darwin, a browser-based labelling tool. The data created in Darwin was exported and saved in the DermX repository.

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

Data was collected through manual human annotation, using a browser-based annotation tool. Dermatologists were allowed a period of getting accustomed to both the annotation protocol and the annotation tool before evaluating DermX.

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

Original images were randomly sampled from the six target conditions, up to a cap of 100 images per condition.

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)?

Board-certified dermatologists annotated the data. They are hired as consultants by Omhu, and their salaries are confidential information.

Over what timeframe was the data collected?

Annotations were collected over the course of two weeks.

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

No.

Does the dataset relate to people?

Yes.

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

Original images were obtained via the two public datasets mentioned above. Annotations were obtained directly from the eight board-certified dermatologists.

Were the individuals in question notified about the data collection?

The annotators were notified, and the owners of the original images gave their consent.

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

Annotators consented for their data to be used. The owners of DermNetNZ and SD-260 consented for their data to be used in this dataset.

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

Being hired as consultants, the annotators' work belongs to Omhu. The owners of the two datasets may request their data to be removed from DermX.

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.

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)?

Images that all labellers tagged with reasons for discard were removed from the dataset.

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

The annotations for the discarded images are available in the dataset.

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

No software was used to clean the labels.

Uses

Has the dataset been used for any tasks already?

No.

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

No.

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

Guided attention training, characteristic segmentation.

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

No.

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.

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

Currently it is stored on GitHub. In the future, it will be available on DTU Data, and will be associated with a DOI.

When will the dataset be distributed?

It 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 licence.

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

The original images remain the property of DermNetNZ and SD-260, and the restrictions on use for each dataset are maintained for these images.

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

No.

Maintenance

Who is supporting/hosting/maintaining the dataset?

The author is supporting and maintaining the dataset. DTU will be hosting it.

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

Either of the two email addresses listed in the paper.

Is there an erratum?

No.

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

The dataset will be updated as needed by the author.

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)?

No.

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

Yes, through versioning systems.

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

The annotation protocol is provided in the paper. New annotators should be able to contribute to the dataset as long as they follow the protocol.