SkinCon 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 created to enable research on interpretability/explainability and error analysis using human-level concepts labeled by domain experts. With that in mind, each image of skin disease was labeled by a dermatologist with the clinical descriptor terms that describe the lesion. These are a set of clinical terms used in clinical training used for describing lesions. This meta-data can be used in turn for developing and testing interpretable/explainable methods (e.g. concept bottleneck models) or for developing and testing methods that assess classification errors.

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 two existing datasets: Fitzpatrick 17k (Groh et al., 2021) and the Diverse Dermatology Images dataset (Daneshjou et al., 2022). Previously, both of these datasets have been used for benchmarking dermatology AI algorithms across diverse skin tones. However, other than skin tone, no other concept-based labels were previously available for either of these datasets.

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

This dataset is targeting any task where human-level concepts could be used for interpretation/explanation or further analysis.

D. Who funded the creation of the dataset?

There was no funding allocated for the creation of the dataset.

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 are images of skin disease that originally include information about disease diagnosis and skin tone. We have densely annotated images with the clinical descriptors used by dermatologists to describe lesions, which encapsulate shape, color, and texture.

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

In Table II, we show the number of concepts represented in each dataset; a single image may be labeled by multiple concepts.

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 is a single clinical image with a diagnosis, whether that diagnosis is a benign or malignant process, skin tone using the Fitzpatrick skin tone scale (I-VI), as well as all the clinical concepts present within that image. Clinical concepts are description terms used by dermatologists for describing a lesion's shape, texture, and color. There is no age and gender information associated with this dataset. In Table I, we provide the distribution of images for each skin tone.

TABLE I

Distribution of images over skin tones. The Fitzpatrick skin tone scale has been used by dermatologists and AI practitioners for assessing skin color. Fitzpatrick I-II represents white skin while Fitzpatrick V-VI represents brown and black skin.

Fitzpatrick Skin Tone	#Images (Fitzpatrick17k)	# Images (DDI)
I-II	1738	208
III-IV	1350	241
V-VI	467	207
Unknown	135	-

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

Yes, the images were previously labeled with skin disease, benign versus malignant, and Fitzpatrick skin type (Groh et al., 2021; Daneshjou et al., 2022). We have added clinical concept labels; the full list can be seen in II. These clinical

TABLE II

CONCEPT STATISTICS FOR THE DATASET. IN TOTAL, WE HAVE $3230 + 656 = 3886$ IMAGES.

Name of the concept	Number of images with concept
Vesicle	46
Papule	1580
Macule	37
Plaque	2135
Abscess	5
Pustule	103
Bulla	64
Patch	155
Nodule	235
Ulcer	167
Crust	550
Erosion	214
Excoriation	46
Atrophy	70
Exudate	157
Purpura/Petechiae	10
Fissure	32
Induration	33
Xerosis	35
Telangiectasia	105
Scale	789
Scar	127
Friable	167
Sclerosis	27
Pedunculated	34
Exophytic/Fungating	50
Warty/Papillomatous	64
Dome-shaped	213
Flat topped	18
Brown(Hyperpigmentation)	1123
Translucent	23
White(Hypopigmentation)	279
Purple	87
Yellow	268
Black	119
Erythema	2374
Comedo	27
Lichenification	26
Blue	5
Umbilicated	54
Poikiloderma	5
Salmon	13
Wheal	21
Acuminate	8
Burrow	5
Gray	5
Pigmented	6
Cyst	6

labels come from the lexicon of terms used by dermatologists to describe skin lesions (Bolognia et al., 2017). Two board-certified dermatologists helped create the concept list based on the most commonly used terms and by consulting (Bolognia et al., 2017).

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.

Gender and age information is missing. These were not available with the original image datasets.

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).

Most dermatology datasets lack diverse skin tones (Daneshjou et al., 2021). For this dataset, we drew from two datasets that had representations of Fitzpatrick I-VI skin tones and across multiple skin diseases (Groh et al., 2021; Daneshjou et al., 2022). While these datasets do not include all possible diagnoses in dermatology, they do encapsulate a diverse representation; the DDI dataset, in particular, included rare diseases (Daneshjou et al., 2022).

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.

We do not prescribe any particular train, validation, test split. However, we note that the disease labels on DDI are all biopsy-proven diagnoses, which is ideal for benchmarking with medical applications.

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

Our dataset is built off of two existing image datasets (Groh et al., 2021; Daneshjou et al., 2022). The skin diagnosis labels for (Groh et al., 2021) likely have label noise; the dataset comes from online dermatology atlases. Per (Groh et al., 2021), only 504 images underwent further diagnosis review with 69% of images being considered diagnostic of the labeled condition. Further details are described in (Groh et al., 2021). In contrast, DDI had every image labeled by biopsy-proven diagnosis (Daneshjou et al., 2022). Our concept labels are provided by dermatologists who are trained to use a clinical lexicon of descriptors for describing lesions. However, this does not mean that there is no label noise with the use of this lexicon; prior research has not looked at agreement among dermatologists in the use of these terms. However, learning to use these terms is a central part of dermatological training, as they are used to provide lesion descriptions both in clinical documentation and oral communication.

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 built off of two existing image datasets (Groh et al., 2021; Daneshjou et al., 2022).

(Groh et al., 2021) developed the Fitzpatrick 17k dataset by pulling images from DermaAmin (AlKattash) and Atlas Dermatologico (da Silva). The Fitzpatrick 17k dataset can be accessed here: `https://github.com/mattgroh/fitzpatrick17k` and is maintained by Matt Groh. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

The DDI dataset was developed and is maintained by Roxana Daneshjou and colleagues (Daneshjou et al., 2022). Data use is governed by the Stanford Research Use Agreement, as well as the Terms of Use of the Stanford University School of Medicine. The website is here: `https://ddi-dataset.github.io/`. The data is hosted by Stanford AIMI: `https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965`.

There is no cost to use the datasets for non-commercial research. Because these are medical images, both datasets require agreeing to data use agreements (such as not attempting to re-identify patients).

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 for collecting images and labeling them with diagnosis and Fitzpatrick skin tone are described previously for Fitzpatrick 17k (Groh et al., 2021) and DDI (Daneshjou et al., 2022).

To choose the concept labels that describe a lesion's color, texture, and shape, we had two board-certified dermatologists review the clinical nomenclature used for describing lesions and select useful terms based on their clinical training (Bolognia et al., 2017). Forty-eight clinical concepts were selected. Then each image was labeled by at least one dermatologist using a labeling tool where each image was presented along with all the possible concepts (multiple concepts could be selected per an image).

A board-certified dermatologist validated the concept annotations. We showed the image, along with the concept annotations, and asked which of the annotations they agreed with. In total, a board-certified dermatologist validated 323 10% of the images; the validator agreed with $1056/1082 = 97.6\%$ of the concept annotations from the Fitzpatrick17k subset.

To get further validation, all of the images from the DDI dataset (656 images) and a random selection of 300 images from the Fitzpatrick dataset were independently labeled using the same labeling interface as was used in the initial labeling procedure, for a total of 956 images. 94% of these images were of sufficient quality across all labelers. Validation labels were provided by two dermatologists with 12 and 5 years of dermatology experience. Overall, we found that independent validators' annotations agreed with SkinCon labels 94% of the time – 92% for Fitzpatrick and 94% for DDI.

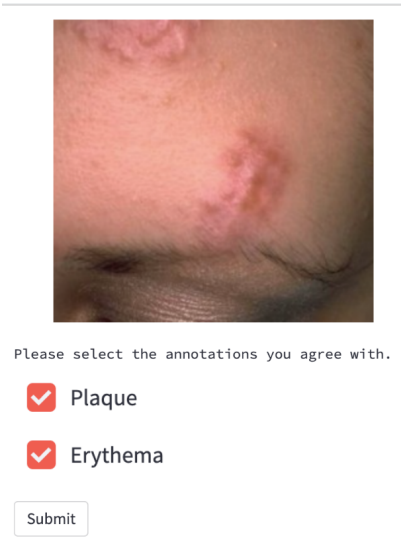

Fig. 1. **An example from the validation interface.**

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.

Skin disease images were reviewed by domain experts (dermatologists) for assessment and labeling.

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

We randomly sampled images without replacement from the Fitzpatrick17k dataset. The entire DDI dataset was used.

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 data was labeled by board-certified dermatologists. They were not compensated. However, as they were involved in study design and conception, labeling, and writing, they are authors of the SkinCon paper.

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.

Per the original papers, the skin images likely came from years of cases (Groh et al., 2021; Daneshjou et al., 2022).

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. Labeling was done by humans as described above.

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?

The processes previously described do achieve the motivation for this dataset.

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 Fitzpatrick 17k dataset can be accessed here: `https://github.com/mattgroh/fitzpatrick17k` and maintained by Matt Groh. This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

The DDI dataset was developed and is maintained by Roxana Daneshjou and colleagues (Daneshjou et al., 2022). Data use is governed by the Stanford Research Use Agreement, as well as to the Terms of Use of the Stanford University School of Medicine. The website is here: `https://ddi-dataset.github.io/`. The data is hosted by Stanford AIMI: `https://stanfordaimi.azurewebsites.net/datasets/35866158-8196-48d8-87bf-50dca81df965`.

An image dump of Fitzpatrick 17k requires filling out a form on the aforementioned github page. However, one could also pull the images directly from DermaAmin (AlKattash) and Atlas Dermatologico (da Silva). Therefore, Fitzpatrick 17k does have redundancy.

DDI is hosted by Stanford AIMI. Since it is de-identified medical images, it does require agreeing to the terms of use and creating a login to access the data.

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

The datasets are already available. Fitzpatrick 17k is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

For DDI, data use is governed by the Stanford Research Use Agreement, as well as to the Terms of Use of the Stanford University School of Medicine.

SkinCon labels are licensed under MIT.

SkinCon labels and information can be found on the SkinCon website `https://skincon-dataset.github.io/`.

C. Are there any copyrights on the data?

See above for licensing for datasets. Note that both Fitzpatrick 17k and DDI have terms around non-commercial use.

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

There are no fees. However, there are restrictions on use because the data include medical images, there are restrictions on data use (for example, no re-identification allowed).

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, the dataset may be expanded, but there is currently no plan for doing so.

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

Updates will be communicated through the SkinCon website `https://skincon-dataset.github.io/`.

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

Through the SkinCon website `https://skincon-dataset.github.io/`.

E. 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?

Currently, we do not have mechanisms in place; however, others may contact us to discuss potential use cases.

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.

The addition of the SkinCon labels did not require IRB approval since we used previously released datasets.

DDI had IRB approval for sharing de-identified images as described in the original paper (Daneshjou et al., 2022).

Fitzpatrick 17k pulled from public internet dermatology atlases (Groh et al., 2021).

B. 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)? If so, please provide a description.

DDI contains deidentified medical images; there are guidelines around the use of this dataset; SkinCon fits within this guidance (Daneshjou et al., 2022). Fitzpatrick 17k pulled from public dermatology atlases (Groh et al., 2021).

The clinical descriptor labels describe what is seen in the 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 has depictions of skin diseases which may be distressing.

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

Yes

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. However, skin tone is labeled and the distribution is shown in Table I.

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.

Images in the Fitzpatrick 17k dataset include identifiable images that are currently available on two online dermatology atlases (da Silva; AlKattash). DDI only includes deidentified images. Both datasets include language around not identifying individuals as part of the terms of use.

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.

The original datasets have skin disease diagnoses.

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

We used existing released datasets.

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 DDI dataset is a retrospective, deidentified dataset developed with IRB approval; de-identified retrospective datasets do not have any connection to the individuals (Daneshjou et al., 2022). The Fitzpatrick 17k dataset was based on publicly available skin atlas images (Groh et al., 2021). The consent process for the dermatology atlases was not described; we contacted the creators of the original dermatology atlases to ascertain the process by which the images were acquired, but did receive a response (AlKattash; da Silva).

We provide dense labels to be used with these existing datasets.

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?

The labels we provide are clinical descriptors of the images describing the color, shape, and texture seen in these images. This information is not sensitive as it does not provide anything beyond what can already be visualized in the image.

VIII. Acknowledgment

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

Jehad Amin AlKattash. Dermaamin. URL `https://www.dermaamin.com/site/`.

Bolognia, Schaffer, and Cerroni. *Dermatology*. Elsevier, 2017. ISBN 9780702062759.

Samuel Freire da Silva. Atlas dermatologico. URL `http://atlasdermatologico.com.br/`.

Roxana Daneshjou, Mary Smith, Mary Sun, Veronica Rotemberg, and James Zou. Disparities in dermatology ai performance on a diverse, curated clinical image set. *JAMA Dermatology*, 157(11), 2021.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *arXiv preprint arXiv:2203.08807*, 2022.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1820–1828, 2021.