ChestX-ray8 Datasheet

Christian Garbin

November 2020

1 ChestX-ray8

This paper converts the description of the ChestX-ray8 dataset [12] from the prose of its original paper into the structured format of a dataset datasheet [2]. It originally had images for eight diseases, enhanced later to cover fourteen diseases, resulting in the other name by which this dataset is known, ChestX-ray14. The paper describing the dataset still refers to it as “ChestX-ray8”.

ChestX-ray8 is a dataset with over 100,000 chest X-ray images and their labels. It was created and made publicly available by the National Institutes of Health Clinical Center [4].

Information for the datasheet was compiled from:

- The latest (fifth) version of the paper [12].
- The NIH news release [4].
- The initial review of the dataset by L. Oakden-Rayner [7].
- The detailed analysis of problems with the dataset by L. Oakden-Rayner [8].
- The study of of different CNNs using ChestX-ray8 [1].
- The detailed description of the instances [5].
- The README file for the dataset [5].
- The FAQ for the dataset [5].

The datasheet is written from the first-person point of view, as if the authors had created it, to make it more realistic. Whenever applicable, the source for the information used in the datasheet is cited.

2 Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Deep learning has shown promising results in some medicals applications. However, methods proposed so far "are evaluated on some small-to-middle scale problems of (at most) several hundred patients. It remains unclear how well the current deep learning techniques will scale up to tens of thousands of patient studies." [12]

"Particularly for chest X-rays, the largest public dataset is OpenI . . . that contains 3,955 radiology reports from the Indiana Network for Patient Care and 7,470 associated chest X-rays from the hospitals picture archiving and communication system (PACS)." [12]

Creating large datasets for medical applications is hindered by the labeling process. Medical images cannot be labeled with the crowdsource methods used for generic image datasets. On the other hand, X-ray images are accompanied by extensive radiological reports. We use natural language processing (NLP) to extract labels from these reports.

With the automated label process in place, we were able to create a dataset with over 100,000 frontal X-ray images, annotated with fourteen common thoracic diseases. We name this dataset "ChestX-ray8" (originally from the eight diseases we were able to mine from the reports, later enhanced to fourteen — we kept the name for historical reasons).

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 a team from the National Institutes of Health Clinical Center.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

"This work was supported by the Intramural Research Program of the NIH Clinical Center (clinicalcenter.nih.gov) and National Library of Medicine (www.nlm.nih.gov). We thank NVIDIA Corporation for the GPU donations." [5]

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a frontal chest X-ray and a set of labels that identifies diseases in the image. One image may have up to fourteen labels, as multiple diseases may be present in one patient. Instances labeled "No finding" could contain disease patterns other than the listed fourteen or uncertain findings within the fourteen categories.

Each instance also includes patient ID, gender, age, follow up number, view position, original image size, and original pixel spacing.

Figure 1 shows the distribution of findings, age groups, gender, view position in the dataset and in the train/test set.

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

There are 112,120 images of 30,805 unique patients [5]. Table 1 shows the number of images with each disease.

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).

The dataset was created from the Clinical Center's PACS system. "Given those 8 text keywords, we search the PACS system to pull out all the related radiological reports (together with images) as our target corpus. A variety of Natural Language Processing (NLP) techniques are adopted for detecting the pathology keywords and removal of negation and uncertainty. Each radiological report will be either linked with one or more keywords or marked with 'Normal' as the background category." [12]

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance is an image extracted from DICOM and downscaled to 1024 × 1024 pixels. Images are stored in PNG format.

Besides the image files, each image has the following data stored in a CSV file, whose key is the image file name:

- File name: the complete name of the file, including the extension. The name of the file encodes the patient ID. For example, the file 00000003_000.png is an image for the patient ID 3.
- Finding labels: a list of diseases labels extracted from the radiological report.
- Follow-up #: the follow-up sequence number.
- Patient ID: a number that uniquely identifies a patient. It is a sequential number that does not encode any other information.
- Patient age: patient age in years at the time the image was taken.
- Patient gender: biological gender (male or female) of the patient.
- View position: either PA (posterioranterior) or AP (anteriorposterior).
- Original image width: the width of the original image in DICOM in pixels.
- Original image height: the height of the original image in DICOM in pixels.
- Original image pixel spacing X: original image column spacing from DICOM.
- Original image pixel spacing Y: original image row spacing from DICOM.

Bounding boxes are available for 983 images. The bounding boxes were created by a board-certified radiologist.

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

Each image has one or more labels indicating a disease. Multiple diseases are separated by |(vertical bar). For example, "Mass|Nodule" indicates that "Mass" and "Nodule" were extracted from the report. Multiple labels are listed in alphabetical order. Instances labeled "No finding" could contain other diseases (other than the listed fourteen) or uncertain findings within the fourteen diseases.

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.

Everything is included in the dataset.

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.

One patient may have multiple images. The "patient ID" column in the dataset description identifies the patient. In addition to that, the "follow-up #" column indicates a follow-up for a patient, thus a chronological order for the images of a patient.

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 comes with a specified train/test split, such that none of the patients in the training split are in the test split and vice versa.

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

The disease labels were extracted from the radiological reports with NLP. While "[t]he text-mined disease labels are expected to have accuracy >90%." [4], "[t]here are several things about the published image labels that we want to clarify:

1. Different terms and phrases might be used for the same finding: The image labels are mined from the radiology reports using NLP techniques. Those disease keywords are purely extracted from the reports. The radiologists often described the findings and impressions by using their own preferred terms and phrases for each particular disease pattern or a group of patterns, where the chance of using all possible terms in the description is small.

2. Which terms should be used: We understand it is hard if not impossible to distinguish certain pathologies solely based on the findings in the images. However, other information from multiple sources may be also available to the radiologists (e.g. reason for exam, patients' previous studies and other clinical information) when he/she reads the study. The diagnostic terms used in the report (like 'pneumonia') come from a decision based on all of the available information, not just the imaging findings.

3. Entity extraction using NLP is not perfect: we try to maximize the recall of finding accurate disease findings by eliminating all possible negations and uncertainties of disease mentions. Terms like 'It is hard to exclude ...' will be treated as uncertainty cases and then the image will be labeled as 'No finding'.

4. 'No finding' is not equal to 'normal'. Images labeled with 'No finding' could contain disease patterns other than the listed 14 or uncertain findings within the 14 categories.

[5] ”

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.

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)? If so, please provide a description.

Although the dataset has images from patients, “[it] was rigorously screened to remove all personally identifiable information before release.” [4]

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

No.

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

Yes. The dataset contains X-rays from human patients.

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.

The dataset identifies the gender and age in years of each patient. Figure 2 shows the distribution of findings across age groups and gender.

There are indications in the literature that gender imbalance in the training dataset affects the model performance [3]. This dataset is well-balanced in overall gender distribution, as well as findings for each gender.

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.

No. “With patient privacy being paramount, the dataset was rigorously screened to remove all personally identifiable information before release." [4]

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 dataset identifies diseases extracted from radiological reports. Other information about the patient's health may be inferred from the images. However, there is no link between an image and the identity of the patient.

Any other comments?

Collection Process

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.

"A variety of Natural Language Processing (NLP) techniques are adopted for detecting the pathology keywords and removal of negation and uncertainty. Each radiological report will be either linked with one or more keywords or marked with 'Normal' as the background category. [E]ach image is labeled with one or multiple pathology keywords or "Normal" otherwise." [12]

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 images were extracted from our institute's DICOM system. The labels were "mined from our institute's PACS system. First, we short-list eight common thoracic pathology keywords that are frequently observed and diagnosed, i.e., Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumathorax ..., based on radiologists' feedback. Given those 8 text keywords, we search the PACS system to pull out all the related radiological reports (together with images) as our target corpus." [12]

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

The patients were selected from the NIH Clinical Center. PACS system records, covering 1992 to 2015. Out of all records, the ones with the eight (later fourteen) disease keywords were selected for the dataset.

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 board-certified radiologist verified the terms used to mine the PACS system. A board-certified radiologist added bounding boxes to selected 983 images, covering 200 instances of each disease.

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.

The data was collected in 2017. The PACS records are from 1992 to 2015.

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.

Unknown.

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

Yes.

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

Data was extracted from the NIH Clinical Center DICOM and PACS systems.

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.

Patients are notified of the X-ray and medical records collection as part of their consultation.

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.

Unknown. However, the NIH Clinical Center *Legal, Ethical, and Safety Issues* document states that "[s]ome of the information obtained from you may appear in scientific publications or be presented to professional audiences at meetings. It may be used for the purpose of teaching health professionals or students in the health professions. Under these circumstances, measures are taken to conceal your identity." [6]

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).

N/A

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.

Unknown. However, the dataset does not have information that directly identifies individual patients.

Any other comments?

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)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

" [1]Overall, our approach produces labels using the reports in two passes. In the first iteration, we detected all the disease concept in the corpus. The main body of each chest X-ray report is generally structured as "Comparison", "Indication", "Findings", and "Impression" sections. Here, we focus on detecting disease concepts in the Findings and Impression sections. If a report contains neither of these two sections, the full-length report will then be considered. In the second pass, we code the reports as "Normal" if they do not contain any diseases (not limited to 8 predefined pathologies).

Pathology Detection: We mine the radiology reports for disease concepts using two tools, DNorm ...and MetaMap .... DNorm is a machine learning method for disease recognition and normalization. It maps every mention of keywords in a report to a unique concept ID in the Systematized Nomenclature of Medicine Clinical Terms (or SNOMED-CT), which is a standardized vocabulary of clinical terminology for the electronic exchange of clinical health information.

MetaMap is another prominent tool to detect bioconcepts from the biomedical text corpus. Different from DNorm, it is an ontology-based approach for the detection of Unified Medical Language System©(UMLS©) Metathesaurus. In this work, we only consider the semantic types of Diseases or Syndromes and Findings (namely

[1] The description of the label extraction, being an innovative part of the work, is one of the strong parts of the original paper [12]. That entire section is quoted here.

'dsyn' and 'fndg' respectively). To maximize the recall of our automatic disease detection, we merge the results of DNorm and MetaMap. . . .

Negation and Uncertainty: The disease detection algorithm locates every keyword mentioned in the radiology report no matter if it is truly present or negated. To eliminate the noisy labeling, we need to rule out those negated pathological statements and, more importantly, uncertain mentions of findings and diseases, e.g., "suggesting obstructive lung disease".

Although many text processing systems . . . can handle the negation/uncertainty detection problem, most of them exploit regular expressions on the text directly. One of the disadvantages to use regular expressions for negation/uncertainty detection is that they cannot capture various syntactic constructions for multiple subjects. For example, in the phrase of "clear of A and B", the regular expression can capture "A" as a negation but not "B", particularly when both "A" and "B" are long and complex noun phrases ("clear of focal airspace disease, pneumothorax, or pleural effusion" in Fig. 3).

To overcome this complication, we handcraft a number of novel rules of negation/uncertainty defined on the syntactic level in this work. More specifically, we utilize the syntactic dependency information because it is close to the semantic relationship between words and thus has become prevalent in biomedical text processing. We defined our rules on the dependency graph, by utilizing the dependency label and direction information between words.

As the first step of preprocessing, we split and tokenize the reports into sentences using NLTK . . . . Next we parse each sentence by the Bllip parser . . . using David McCloskys biomedical model . . . . The syntactic dependencies are then obtained from "CCProcessed" dependencies output by applying Stanford dependencies converter . . . on the parse tree. The "CCProcessed" representation propagates conjunct dependencies thus simplifies coordinations. As a result, we can use fewer rules to match more complex constructions. For an example as shown in Fig. 3, we could use "clear → prep of → DISEASE" to detect three negations from the text (neg, focal airspace disease), (neg, pneumothorax), and (neg, pleural effusion).

Furthermore, we label a radiology report as "normal" if it meets one of the following criteria:

- If there is no disease detected in the report. Note that here we not only consider 8 diseases of interest in this paper but all diseases detected in the reports.
- If the report contains text-mined concepts of "normal" or "normal size" (CUIs C0205307 and C0332506 in the SNOMED-CT concepts respectively).

" [12]

Images: The images were extracted from DICOM and resized to 1024 × 1024 pixels (from the usual 3000 × 2000 pixels of an X-ray image). "Their intensity ranges are rescaled using the default window settings stored in the DICOM header files." [12] The resulting grayscale is 0-255 [9].

Bounding boxes: A smaller set of images has a bounding box for the diseases. "[W]e first select 200 instances for each pathology (1,600 instances total), consisting of 983 images. Given an image and a disease keyword, a board-certified radiologist identified only the corresponding disease instance in the image and labeled it with a B-Box. The B-Box is then outputted as an XML file." [12]

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.

The raw data is available in the DICOM and PACS system. However, it will not be made available to the public because it contains personally identifiable data and other private pieces of information.

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

No

Any other comments?

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Together with the dataset “we present a unified weakly-supervised multi-label image classification and pathology localization framework, which can detect the presence of multiple pathologies and subsequently generate bounding boxes around the corresponding pathologies.” [12]

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

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

“By using this free dataset, the hope is that academic and research institutions across the country will be able to teach a computer to read and process extremely large amounts of scans, to confirm the results radiologists have found and potentially identify other findings that may have been overlooked.” [4]

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The image downsampling and reduction in gray levels to 256 may result in missing some important artifacts. “Of note, subtle pneumothoraces, small nodules, and retrocardiac opacities become nearly impossible to diagnose for a human expert.” [9]

The images were collected at one institution only. It is known that the devices and methods used in an institution may result in markers that a neural network may learn, instead of learning the disease-specific markers [14] [10] [13].

“The image labels are NLP extracted so there would be some erroneous labels but the NLP labelling accuracy is estimated to be >90%.” [5]. However, “[the label problems] mean the dataset as defined currently is not fit for training medical systems, and research on the dataset cannot generate valid medical claims without significant additional justification.” [8]

Some of the labels, such as *effusion*, *pneumothorax*, and *fibrosis*, need further interpretation and verification [8]. For example, “*pneumothorax* is often labeled for already treated cases (i.e. a drain is visible in the image which is used to treat the pneumothorax) in the ChestX-ray14 dataset. . . . [A network may learn] not only to detect an acute pneumothorax but also the presence of chest drains.” [1] as shown in figure 4. Labels may come not only from the images, but also from “multipl[e] sources [available] to the radiologists (e.g. reason for exam, patients’ previous studies and other clinical information) when he/she reads the study.” [5]

This is due to the nature of radiology reports. “Radiology reports are not objective, factual descriptions of images. The goal of a radiology report is to provide useful, actionable information to their referrer, usually another doctor.” [8]

Are there tasks for which the dataset should not be used? If so, please provide a description.

We do not recommend using this dataset alone to develop clinical applications. Please review the previous item for details.

Any other comments?

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? If so, please provide a description.

Yes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?

The dataset is available in the NIH Clinical Center Box folder at `https://nihcc.app.box.com/v/ChestXray-NIHCC`.

When will the dataset be distributed?

The dataset is already available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or un-

der applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

"The usage of the data set is unrestricted. But you should provide the link to our original download site, acknowledge the NIH Clinical Center and provide a citation to our CVPR 2017 paper." [5]

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown.

Any other comments?

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The NIH Clinical Center hosts and maintains the dataset.

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

Refer to the README file in the dataset folder.

Is there an erratum? If so, please provide a link or other access point.

The dataset folder has a log file that list all changes, including corrections.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Updates are done as needed and noted in the log file. There is no notification mechanism. Users of the dataset are requested to check the log file periodically.

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)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Older versions are not maintained.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No.

Any other comments?

Finding	Male	Female	Total	%
No finding	33,922	26,439	60,361	53.8
Infiltration	5,383	4,164	9,547	8.5
Atelectasis	2,603	1,612	4,215	3.8
Effusion	2,158	1,797	3,955	3.5
Nodule	1,524	1,181	2,705	2.4

Pneumothorax	1,001	1,193	2,194	2.0
Mass	1,301	838	2,139	1.9
Effusion + Infiltration	942	661	1,603	1.4
Atelectasis + Infiltration	844	506	1,350	1.2
Consolidation	771	539	1,310	1.2
Atelectasis + Effusion	680	485	1,165	1.0
Pleural_Thickening	660	466	1,126	1.0
Cardiomegaly	508	585	1,093	1.0
Emphysema	562	330	892	0.8
Infiltration + Nodule	492	337	829	0.7
Atelectasis + Effusion + Infiltration	424	313	737	0.7
Fibrosis	387	340	727	0.6
Edema	329	299	628	0.6
Cardiomegaly + Effusion	209	275	484	0.4
Consolidation + Infiltration	243	198	441	0.4
Infiltration + Mass	263	157	420	0.4
Effusion + Pneumothorax	202	201	403	0.4
Effusion + Mass	229	173	402	0.4
Atelectasis + Consolidation	212	186	398	0.4
Mass + Nodule	246	148	394	0.4
Edema + Infiltration	193	199	392	0.3
Infiltration + Pneumothorax	188	157	345	0.3
Emphysema + Pneumothorax	202	135	337	0.3
Consolidation + Effusion	202	135	337	0.3
Pneumonia	194	128	322	0.3
All other findings	6,266	4,603	10,869	9.7

Table 1: ChestX-ray8 top thirty findings. Multiple findings in one image are listed in alphabetical order. "No finding" is not equal to "normal". Images labeled with "No finding" could contain disease patterns other than the listed fourteen or uncertain findings within the fourteen categories [5].

3 Appendix

3.1 Datasheets for datasets

Datasheets for Datasets "document [the dataset] motivation, composition, collection process, recommended uses, and so on. [They] have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted biases in machine learning systems, facilitate greater reproducibility of machine learning results, and help researchers and practitioners select more appropriate datasets for their chosen tasks."

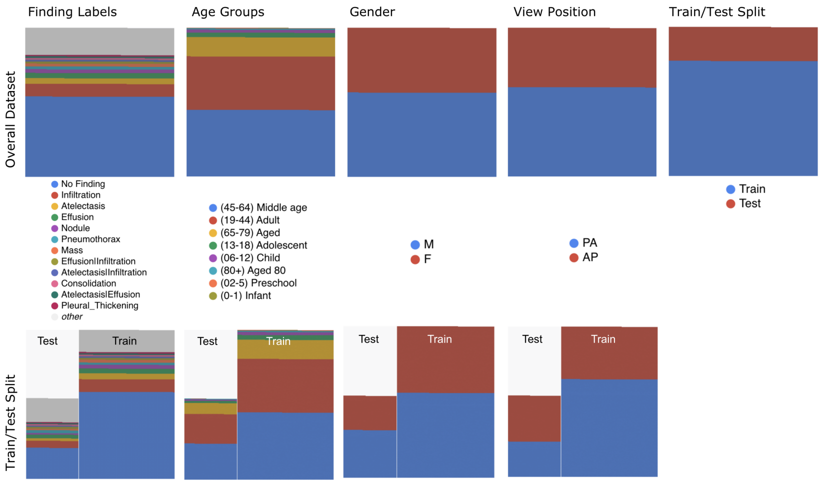

Figure 1: ChestX-ray8 distribution of images by findings, age groups, gender, and view position. The first row shows the overall dataset. The second rows shows the split for the train/test sets. In general, the train/test split is well balanced across category. There is a slightly larger proportion of the "AP" view position in the test set[a].

[a] Visualizations created with Google Facets and available in this GitHub repository.

The motivation behind the proposal was the electronics industry, where every component has a datasheet that describes its operating characteristics and recommended uses. In machine learning, data is the input for model training. Using the wrong dataset, or using a dataset outside of its original intent, or even not understanding well enough the limitations of a dataset, has dire consequences for the model. However, "[d]espite the importance of data to machine learning, there is no standardized process for documenting machine learning datasets. To address this gap, we propose datasheets for datasets."

References

[1] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach. Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. *Scientific Reports*, 9(1), apr 2019.

[2] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford. Datasheets for Datasets v7. Link to publication 2020-08-06, 2018.

[3] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, may 2020.

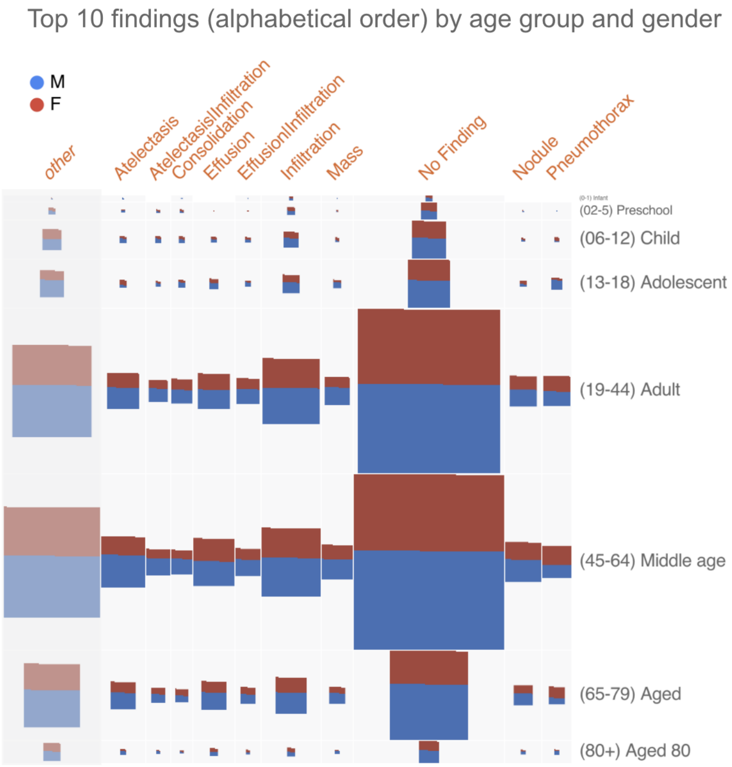

Figure 2: ChestX-ray8 top ten findings by age group and gender (and "other" to account for the other findings). Genders are well balanced across age groups. Adults and middle ages are the larger groups. Younger and older ages have fewer images[a].

[a] Visualizations created with Google Facets and available in this GitHub repository.

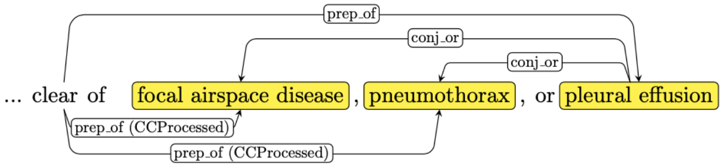

Figure 3: ChestX-ray8 text parser dependency graph to detect three negations in the text: "clear of focal airspace disease, pneumothorax, or pleural effusion". [12]

[4] National Insititutes of Health. NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community. Link to publication 2020-07-27, 2017.

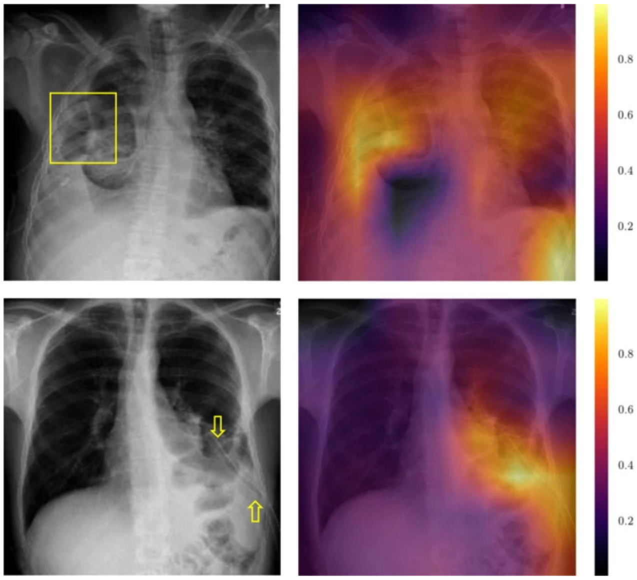

Figure 4: Both images are labeled with “pneumothorax”. In the top image the patient has not been treated yet. In the bottom image the patient has been treated (chest drains indicated with arrows). Using Grad-CAM [11] to visualize the important areas for prediction shows that models incorrectly learn to predict based on the treatment, not the pathology (bottom image) [1].

[5] National Institutes of Health Clinical Center. ChestX-ray8. Link to publication 2020-08-05, 2017.

[6] National Institutes of Health Clinical Center. Legal, Ethical, and Safety Issues. Link to publication 2020-08-05, 2017.

[7] L. Oakden-Rayner. Quick thoughts on ChestXray14, performance claims, and clinical tasks. Link to publication 2020-07-27, 2017.

[8] L. Oakden-Rayner. Exploring the ChestXray14 dataset: problems. Link to publication 2020-07-27, 2018.

[9] L. Oakden-Rayner. Half a million x-rays! First impressions of the Stanford and MIT chest x-ray datasets. Link to publication 2020-07-27, 2019.

[10] E. H. P. Pooch, P. L. Ballester, and R. C. Barros. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. 2019.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-

Based Localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.

[12] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases v5. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017.

[13] L. Yao, J. Prosky, B. Covington, and K. Lyman. A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging. *arXiv e-prints*, page arXiv:1904.01638, Apr. 2019.

[14] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, nov 2018.