A Appendix

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.

Indian-CT was constructed to provide chest CT scans of COVID-19 patients from India for building machine learning models that can aid in analyzing lungs for the diagnosis of the disease. This dataset can also be used as a category while detecting other pulmonary diseases. To our knowledge, this is the only chest CT-scan dataset from the Indian population. Since all the images included in this dataset are from a single hospital and acquired using a single CT scanner, any confounding factors in the data are minimized.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The research team, including the authors, Suba S, Nita Parekh, Ramesh Loganathan, Vikram Pudi, from International Institute of Information Technology, Hyderabad, collected the data from Gandhi Hospital, Hyderabad.

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 funded by the RAKSHAK (Remedial Action, Knowledge Skimming and Holistic Analysis of COVID-19) project of Department of Science and Technology (DST), India.

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 in the dataset is a slice of CT scan of the lungs taken from COVID-19 patients. Multiple slices from each patient are available and each slice is stored in PNG format. Except the last two digits, the name of the file is unique to a patient.

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

Currently, there are 6174 images taken from 142 COVID-19 patients. Each patient has 40 slices (images) in the dataset.

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).

This dataset does not contain all possible instances. Based on the availability of the data from the hospital, currently, only 142 patient data is available. We will add more eventually.Each CT volume was converted to png format after selecting only slices in the range 40 - 300 as this range was found to consist of the broadest lung window devoid of other internal organs. This heuristic could be applied on all the images as these are obtained from a single CT scanner machine. Every 3rd slice from the chosen range was considered for analysis to reduce the size of the dataset. For a few samples (< 10), however, since sufficient number of slices were not available, every slice in the corresponding range was taken.

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 consists of a slice from the CT scan of the lungs of COVID-19 patients in PNG format. The size of each image is 512 by 512 pixels. The raw data from the hospital was in .dcm format, which was converted to PNG.

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

The dataset consists of only images of COVID-19 class, and hence no separate label is associated.

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

The slices in their raw form consisted of patients' personal identification information such as name, age, date of birth, etc., along with the scanner settings which were removed by default when converting to PNG format.

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.

There are no known relationships between individuals except for the fact that they all are COVID-19 patients.

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 does not contain any splits. We recommend using this dataset as a test set for understanding the generalizability of machine learning models. This could also be used for a detailed analysis of the lungs of COVID-19 patients from the Indian population.

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

None to the best of our knowledge

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 data is self-contained.

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

No. All personal identification details have been removed from the data.

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. The dataset consists of only CT scan images of lungs.

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

Yes.

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 consists of CT scan images of the lungs of COVID-19 patients, collected from a hospital in Hyderabad, India. The age of the patients is in the range of 17 to 79 years, with a mean age of 48 years. All the images were acquired using a single CT scanner, and hence the scanner settings are identical for all images.

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. It is not possible to identify individuals directly or indirectly.

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

No personal data of the patients whose CT scan images is included in available to the users.

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.

The dataset was acquired from the radiology department of Gandhi hospital, Hyderabad. The third-party who assisted in collecting the data named the data COVID-19 CT data collected from the hospital.

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 raw data was acquired using a single CT scanner machine, the details of which are given in Table 1. The images are plain CT scans captured with no contrast and slice thickness of the images are 0.6, 1.5 and 5 mm.

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

A total of 533 patient data was obtained from Gandhi hospital. Of this, 255 patient data were considered for this study. On initial screening of these 255, 108 patient data were removed as these did not exclusively belong to chest CT, had missing information like SliceLocation, or came from different CT scanners. The remaining 278 patient data is under the preprocessing stage and will eventually be added to the Indian-CT dataset. Each CT volume was converted to png format after selecting only slices in the range 40 - 300 as this range was found to consist of the broadest lung window devoid of other internal organs. This heuristic was considered because the CT images in this dataset were obtained using only a single CT scanner machine. Every 3rd slice from the chosen range was considered to reduce the size of the dataset. However, for some samples sufficient slices were not available and in such cases every slice in the corresponding range was taken.

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 third party was involved in the collection process and were funded through RAKSHAK project of Department of Science and Technology (DST), India.

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.

Indian-CT data is collected from Gandhi Hospital, Hyderabad, India from the COVID-19 isolated patients during the period April - September, 2020.

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 study was approved by the institutional review board (IRB) of Gandhi Hospital, Hyderabad. The institutional review boards waived the requirement to obtain written informed consent for this retrospective study. For confidentiality, no link between the patients and the researchers was made available.

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)?

This is a retrospective study, and the data was collected via a third party, from the hospital.

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 study was approved by the institutional review board (IRB) of Gandhi Hospital, Hyderabad. The institutional review boards waived the requirement to obtain written informed consent for this retrospective study. For confidentiality, no link between the patients and the researchers was made available

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.

This is a retrospective study, approved by the institutional review board (IRB) of Gandhi Hospital, Hyderabad. The institutional review boards waived the requirement to obtain written informed consent.

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

There is no link connecting the dataset to the data subjects and hence no potential impact.

Any other comments?

PREPROCESSING/CLEANING/LABELLING

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.

Yes. A total of 533 patient data was obtained from Gandhi hospital. Of this, 255 patient data were considered for this study. On initial screening of the 255 samples, 108 patient data were removed as these did not exclusively belong to chest CT, or had missing information like SliceLocation, or came from different CT scanner. Each CT volume was converted to png format after selecting only slices in the range 40 - 300 as this range was found to consist of the broadest lung window devoid of other internal organs. This heuristic was considered only because the CT images in this dataset was obtained using only a singly CT scanner machine. Every 3rd slice from the chosen range was considered for analysis to reduce the size of the dataset. However, for few samples sufficient slices were not available and in such cases every slice in the corresponding range was taken.

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.

Yes. The raw data is saved but will only be shared at a later stage.

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

Preprocessing was done using a python code.

Any other comments?

USES

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

The dataset has been used as a test set for analyzing the performance of deep learning models on this dataset. The details are given in the paper.

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?

The dataset is the only one of its kind dataset from the Indian population. Apart from using the images for COVID-19 diagnosis problem, this dataset can be used for the analysis of the lungs of patients belonging to the Indian population and can be used as a class in lung disease classification problems. This can be used to test the generalizability of models trained on a different population, as done in this study.

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?

NO

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

None to the best of our knowledge.

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.

The access will be only provided on request for research purposes, after completing an enrollment form in which requesters agree to the collaborative effort for confidential use of the data, and to cite the dataset in their publications. Ethics agreements have been procured for this purpose.

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

Details of how to access the dataset are currently available on the AIMed website (http://aimedhub.iiit.ac.in/). Eventually, this access will be moved to a Data Foundation platform, which is currently being built at IIIT Hyderabad to host several datasets across verticals including health, mobility, smart-cities, etc., as part of a national initiative (details at https://ihub-data.iiit.ac.in/).

When will the dataset be distributed?

The dataset is available upon request.

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

Access to the dataset will be made available on request for research purposes, after completing an enrollment form in which requesters agree to the collaborative effort for confidential use of the data, and to cite the dataset in their publications. Ethics agreements have been procured for this purpose.

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.

Current ethics agreements have authorized the dataset to be used by researchers and collaborators of IIIT Hyderabad and Gandhi Hospital, Hyderabad. The ethics agreements have not authorized the dataset to be hosted directly on any public platform. Thus, it may be made available on request to researchers who agree to the collaborative effort for confidential use of the data.

Any other comments?

MAINTENANCE

Who will be supporting/hosting/maintaining the dataset?

IIIT, Hyderabad will be hosting and maintaining the dataset.

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

By filling the download request form on the website http://aimedhub.iiit.ac.in/.

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

No.

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)?

The dataset will be updated and the updates will be communicated to collaborators via email.

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. There is no applicable limits on retention of the data.

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.

Any changes will be communicated to the collaborators via email.

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.

The ethics agreement in place has not authorized the dataset to be hosted openly on any public platform, and hence the dataset is not available for extension and redistribution. It is still possible for other datasets to provide references to this dataset, but not include this dataset's records directly.

Any other comments?

Other Details

1. Manufacturer details of the CT machine using which the Indian-CT data was collected is given in Table 1.
2. The training of the model was done using an existing dataset, COVIDx-CT-2. The COVIDx CT-2 dataset is released under a CC BY-NC-SA 4.0 license in accordance with the licenses of its constituent datasets. The details of each individual licence and the link from which the dataset was downloaded is: https://www.kaggle.com/hgunraj/covidxct. The new asset that is used for testing is proprietary and is not released yet.

Table 1: Manufacturer details of the CT machine

Key	Value
Manufacturer	SIEMENS
Modality	CT
Manufacturer's Model Name	Emotion 16
Device Serial Number	39306
Software Version(s)	Syngo CT 2014A

3. The study was approved by the institutional review board (IRB) of Gandhi Hospital, Hyderabad. The institutional review boards waived the requirement to obtain written informed consent for this retrospective study., For confidentiality, no link between the patients and the researchers was made available.
4. This study was retrospective and the IRB has waived the requirement of consent.
5. The data was checked manually, sample by sample for any identifiable information and only dummy name, age and dummy date of births were found without any link to address or any other information so that the patient can be traced.
6. This was a retrospective study and hence no potential participant risks are involved and no participant compensation is applicable.