COVIDx-US: An open benchmark ultrasound imaging dataset for AI-driven COVID-19 diagnosis

Ashkan Ebadi, Pengcheng Xi, Stéphane Tremblay
Digital Technologies Research Centre
National Research Council Canada
ashkan,ebadi@nrc-cnrc.gc.ca

Sonny Kohli
Oakville Trafalgar Memorial Hospital
McMaster University, Canada

Alexander MacLean, Alexander Wong
Department of Systems Design Engineering
University of Waterloo, Canada

ABSTRACT

The COVID-19 pandemic continues to have a devastating effect on the health and well-being of the global population. Apart from the global health crises, the pandemic has also caused significant economic and financial difficulties and socio-physiological implications. Effective screening, triage, treatment planning, and prognostication of outcome plays a key role in controlling the pandemic. Recent studies have highlighted the role of point-of-care ultrasound imaging for COVID-19 screening and prognosis, particularly given that it is non-invasive, globally available, and easy-to-sanitize. Motivated by these attributes and the promise of artificial intelligence tools to aid clinicians, we introduce **COVIDx-US**, an open-access benchmark dataset of COVID-19 related ultrasound imaging data. The COVIDx-US dataset was curated from multiple sources and its current version, i.e., v1.2., consists of **150** lung ultrasound videos and **12,943** processed images of patients infected with COVID-19 infection, non-COVID-19 infection, other lung diseases/conditions, as well as normal control cases. The COVIDx-US is the largest open-access fully-curated dataset of its kind that has been systematically curated, processed, and validated specifically for the purpose of building and evaluating artificial intelligence algorithms and models.

INTRODUCTION

The novel Coronavirus Disease 2019 (COVID-19) led to a pandemic of severe and deadly respiratory illness, affecting human lives and well-being. The rapid growth of confirmed cases over several waves of a pandemic highlights the importance of effective screening and risk stratification of infected patients as a means to minimize spread and identify those that need a higher level of care. The reverse transcription-polymerase chain reaction (RT-PCR) test, performed on biological samples taken from the patient, is the main screening method used for COVID-19 detection. Although RT-PCR is used in many countries, it requires a long complicated manual processing that is a huge disadvantage for an effective fight against the pandemic. Moreover, there is no consensus about the sensitivity of RT-PCR testing.

Radiography is an alternative imaging method utilized for COVID-19 screening and risk stratification. As an established method for monitoring and detecting pneumonia, lung point-of-care ultrasound (POCUS) is an emerging imaging modality that is receiving growing attention from the scientific community in recent years. Changes in lung structure, such as pleural and interstitial thickening, are identifiable on lung ultrasound (LUS) and help to detect viral pulmonary infection in the early stages. For COVID-19 screening, recent studies reported identifiable lesions in the bilateral lower lobes and abnormalities in bilateral B-lines on LUS as the main attributes of the disease. Motivated by recent open-source efforts of the research community in the fight against COVID-19 and to support alternative screening, risk stratification, and treatment planning solutions powered by AI and advanced analytics, we introduce **COVIDx-US**, an open-access benchmark dataset of ultrasound imaging data that was carefully curated from multiple sources and integrated systematically specifically for facilitating the building and evaluation of AI-driven analytics algorithms and models.

The current version of the COVIDx-US dataset comprises **150** videos and **12,943** processed ultrasound images of patients diagnosed with COVID-19 infection, non-COVID-19 infection, other lung diseases/conditions, as well as normal control patients. The COVIDx-US dataset was released as part of a large open-source initiative, the COVID-Net initiative, and will be continuously growing, as more data sources become available. To the best of the authors' knowledge, COVIDx-US is the first and largest open-access fully-curated benchmark LUS imaging dataset that is reproducible, easy-to-use, and easy-to-scale thanks to the modular well-documented design.

METHODS

The COVIDx-US dataset continues to grow as new POCUS imaging data is continuously curated and added as part of the broader initiative. All versions of the dataset will be made publicly available. Although this study represents the current snapshot of the dataset in terms of coverage, all the steps, including the data collection and processing pipeline that are introduced in this section in detail, will remain similar in the upcoming versions. **Figure 1** shows the flow of processes and the steps taken to generate the COVIDx-US dataset.

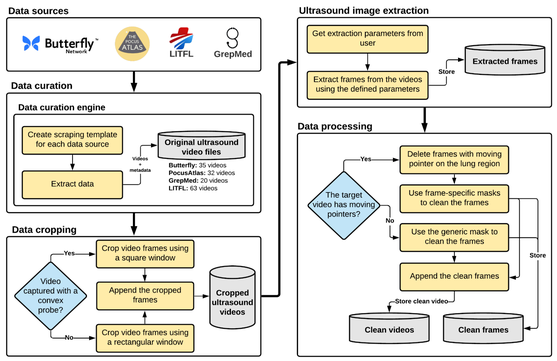

Figure 1. The conceptual flow of COVIDx-US data set integration

DATA SOURCES

The COVIDx-US dataset is heterogeneous in nature, containing ultrasound imaging data of various characteristics, e.g., convex and linear US probes, from multiple sources. The current version, i.e. COVIDx-US v1.2., curates ultrasound video data of four categories, i.e., COVID-19 infection, non-COVID-19 infection (e.g., bacterial infection, non-SARS-CoV-2 viral infection, etc.), other lung diseases/conditions, and normal control, from four different sources: **1)** The POCUS Atlas (TPA), **2)** GrepMed (GM), **3)** Butterfly Network (BN), and **4)** Life in the Fast Lane (LITFL). The COVID-19 US video files account for 39% of the data, although the pandemic is recent. **Table 1** shows the distribution of the LUS video files per data source in the current version of the dataset.

Table 1. 2. Distribution of the collected ultrasound video files per source and class

Data source	Website	Categories				Total
		COVID-19	Non-COVID-19	Normal	Other	
TPA	www.thepocusatlas.com	18	9	5	0	32
GM	www.grepmed.com	8	9	3	0	20
BN	www.butterflynetwork.com	33	0	2	0	35
LITFL	www.litfl.com	0	19	3	41	63
	Total	59	37	13	41	150

Figure 2 shows sample ultrasound frames captured from the ultrasound video recordings in the COVIDx-US dataset. The examples are processed by the COVIDx-US scripts. These few examples illustrate the diversity of ultrasound imaging data in the dataset. The choice of the four different data sources and the heterogeneity in the structure and format of their hosted videos resulted in a highly diverse set of videos and images in the COVIDx-US dataset that is key to the generalizability of the AI-driven solutions that are built on the COVIDx-US dataset.

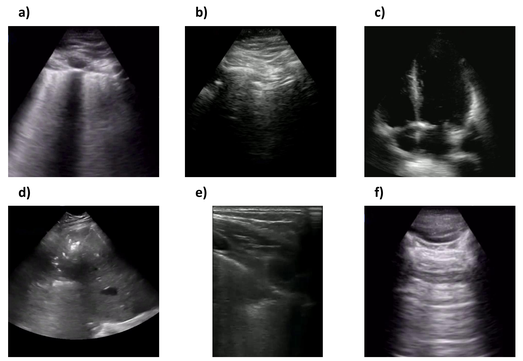

Figure 2. Sample ultrasound frames captured from the curated ultrasound video recordings

DATA CURATION

The data were curated from four data sources, each with a different structure. To support reproducibility and ease of use, we developed data curation engines, personalized for each of the target data sources, to automatically curate lung POCUS video recordings as well as associated metadata from the target data sources and to integrate them locally in a unified, organized structure. No original data is hosted in the COVIDx-US repository and the data is rather curated and integrated locally via our publicly released COVIDx-US scripts and the parameters set by the user.

DATA CROPPING

We treated convex and linear US video files separately. For the convex and linear US video files, we used square and rectangular windows to crop the frames, respectively. We used rectangular windows for linear US video files to include a larger portion of the original file in the processed video file. Publicly available processing scripts that we release as part of COVIDx-US to automatically perform data cropping on the benchmark dataset.

ULTRASOUND IMAGE EXTRACTION

To ensure maximum flexibility of the COVIDx-US dataset and as part of each release, we provide end users with highly flexible data processing scripts, allowing them to extract frames from the initially processed video files based on their research objectives and requirements.

DATA PROCESSING

The final processing step is performed on the extracted frames from the video files as follows: **1)** videos with moving pointers are identified, **2)** if the video contains a moving pointer, frames with a moving pointer on the lung region are deleted and frame-specific masks are generated and stored for the remaining frames, and **3)** if the video does not contain a moving pointer, a generic mask that is suitable for all the extracted frames is generated and stored, and is used to remove artifacts from the frames. All the generated masks are provided as part of the COVIDx-US release. We used an inpainting technique to remove the peripheral artifacts from the frames by replacing bad marks, i.e., pixels in the masked regions, with their neighboring pixels. Clean video files are then generated from the clean frames and both are stored locally on the user's device. **Figure 3** shows a sample cropped ultrasound frame, the mask generated for this specific frame, and the final clean frame obtained by applying the mask to the original frame.

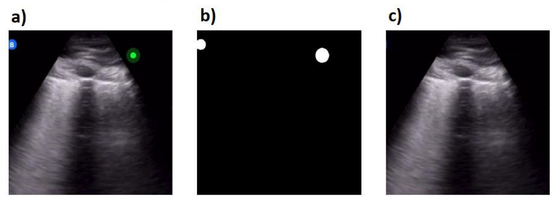

Figure 3. Sample images in COVIDx-US dataset
a) A sample cropped frame, b) the respective generated mask, and c) the resulted clean frame

DATA QUALITY VALIDATION

As a crucial step to ensure the quality of images in the COVIDx-US dataset, our contributing clinician reviewed a randomly selected set of images. His findings and observations confirmed the existence of identifiers and indicators of disease in the COVIDx-US dataset.

CONCLUSION

The COVIDx-US, to the best of our knowledge, is the largest open-access fully-curated benchmark LUS dataset of its kind that is systematically curated, highly reproducible, easy to use, and highly scalable. All the scripts are well-documented and modularly designed to ensure readability and scalability. The scripts, metadata, and generated masks, necessary to reproduce the COVIDx-US data set are available to the general public at NRC-COVID-US GitHub repository. The COVIDx-US dataset will be continuously growing as more data become available. We recommend that users check the COVIDx-US repository frequently, for the latest version of data and scripts.