The Open Kidney Ultrasound Data Set

Rohit Singla rsingla@ece.ubc.ca
Biomedical Engineering, University of British Columbia, Vancouver, BC, Canada

Cailin Ringstrom
Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

Grace Hu
Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

Victoria Lessoway
Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

Janice Reid
Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada

Christopher Nguan
Urologic Sciences, University of British Columbia, Vancouver, BC, Canada

Robert Rohling
Electrical and Computer Engineering and Mechanical Engineering, University of British Columbia, Vancouver, BC, Canada

Appendix A. Datasheet for the Data Set

In this appendix, we provide a completed data sheet for the proposed data set in efforts to increase transparency and accountability.

For what purpose was the data set created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
For the purposes of multi-class kidney segmentation in ultrasound imaging. It focuses on binary segmentation of the capsule from the foreground as well as multi-class segmentation of the cortex, medulla, and central echogenic complex. The literature lacks a standard data set for fair comparisons across algorithms.

Who created this data set (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The initial version was created by Rohit Singla, Cailin Ringstrom, Grace Hu, Victoria Lessoway, Janice Reid, Christopher Nguan and Robert Rohling. They are researchers at the University of British Columbia in Vancouver, British Columbia, Canada.

What support was needed to make this data set? (e.g., who funded the creation of the data set? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
No specific funding or grant was provided for the creation of this data set. The creators are supported by funding from the Natural Sciences and Engineering Council of Canada, the

Kidney Foundation of Canada, and the American Society of Transplant Surgeons.

Any other comments?
n/a.

What do the instances that comprise the data set 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 an ultrasound image with a label for quality, view type, kidney type, and polygon annotations for each class. There are between 0-4 annotations for any given image.

How many instances are there in total (of each type, if appropriate)?
There are a 514 unique B-mode images with 20 additional copies (two sets of 10) repeated from these 514. 534 images in total.

Does the data set contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the data set 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 data set does not contain all possible instances. It is a random sampling of retrospective data from a local institution, rather than a comprehensive global sampling. The sampling is representative of the Canadian population, specifically in the province of British Columbia.

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 ultrasound image in PNG format. The resolution varies depending on the type of machine originally used to acquire the image. Images are randomly sampled from the original video (cine). Images have personally identifying information removed.

Is there a label or target associated with each instance? If so, please provide a description.
Each image is accompanied by a label indicating quality, view type, kidney type, as well as a set of coordinates reflecting the class and polygon annotation of that class.

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.
For confidentiality and privacy reasons, personally identifiable information has been removed from the images. The images were originally stored in the hospital system in DICOM format. Conversion to PNG as well as redaction of information in the images themselves

has been performed.

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 instances.

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 data splits used for training, validation and testing are reported.

Are there any errors, sources of noise, or redundancies in the data set? If so, please provide a description.
A list of errors is maintained online at `https://rsingla92.github.io/kidneyUS/`

There is no existing robustly validated definition of the central echogenic complex. Our definition broadly incorporates different minor parts of the kidney's anatomy.

While the expert sonographers are experts in ultrasound, there many still exist discrepancies between their annotations and how another expert may interpret the image. This aleatoric uncertainty is captured to a small degree in the inter-rater variance above.

Potential sources of error may include inaccurate segmentation masks or incorrect class labels. This has been mitigated through the quality assurance measures performed.

Is the data set self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other data sets)? 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 data set (i.e., including the external resources as they existed at the time the data set 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.
Self-contained.

Does the data set 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.
The nature of medical imaging data including ultrasound is inherently confidential. As a result, we have taken steps to anonymize the images and consulted our institution's ethics board for approval as well as the data release management office.

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

No.

Does the data set relate to people? If not, you may skip the remaining questions in this section
Yes. The data set contains images of people's kidneys.

Does the data set 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 data set
No. The subpopulation of native kidney versus transplant kidney may be the only potential subpopulation that arises from this data.

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 data set? If so, please describe how.
No.

Does the data set 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 data set contains medical imaging data that may or may not include pathology. The images may also be used to ascertain biometrics such as kidney length or width.

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.
After ethics board approval, the data was retrieved from our institution's picture archiving and communications system. The data was then anonymized internally and processed to remove any potentially identifying information. The data was then reviewed by expert sonographers who generated the corresponding labels and annotations.

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?
Original image acquisition was performed by a variety of ultrasound machines and manufacturers. Image retrieval was performed using our institution's picture archiving and communications system. Image anonymization was performed internally using automatic methods. Image annotation and review was performed using the VGG Image Annotator tool.

If the data set is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Random sampling of retrospectively collected ultrasound images. The larger set includes all adult individuals who have received a kidney ultrasound for suspicion of kidney disease, as well as adult kidney transplant recipients who have received an ultrasound. From the larger set, a complete random sampling was taken. One frame per individual is used.

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?
Research assistants, surgeons, and sonographers were involved. No funding was provided.

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.
2014 to 2019

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.
Yes. H21-02375

Does the data set 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)?
The data was retrieved from our institution's storage.

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.
No.

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
No. In our research ethics application, given the scale of the initial retrieval and the subse-

quent anonymization, individuals were not directly notified.

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 data set 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.
No formal analysis has been conducted. Our local data release management office which handles privacy and confidentiality concerns was consulted during our ethics approval process.

Was any pre-processing/cleaning/labelling 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.
Cleaning of the data was performed. DICOM to PNG conversion, meta-data/header cleaning, as well as ultrasound machine-specific removal of personally identifiable information was performed. No additional resizing or processing was performed.

Was the "raw" data saved in addition to the pre-processed/cleaned/labelled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
No

Is the software used to pre-process/clean/label the instances available? If so, please provide a link or other access point.
All software used to process the data is available and open sourced.

Any other comments?
n/a

Has the data set been used for any tasks already?
A trained segmentation model for the kidney capsule task has been additionally used to automatically measure biometrics of kidney length and width in another set of images.

Is there a repository that links to any or all papers or systems that use the data set? If so, please provide a link or other access point `https://rsingla92.github.io/kidneyUS/`

What (other) tasks could the data set be used for?
Automatic biometry, image quality assessment, artefact detection

Is there anything about the composition of the data set or the way it was collected and pre-processed/cleaned/labelled 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?
To the best of our knowledge, there is nothing regarding the composition that may impact future uses.

Are there tasks for which the data set should not be used? If so, please provide a description.
To the best of our knowledge, there are no such tasks.

Will the data set be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the data set was created? If so, please provide a description.
Third parties outside of the creators will be permitted to access the data. This requires registration via the website, including a research use agreement.

How will the data set be distributed (e.g., tarball on website, API, GitHub)? Does the data set have a digital object identifier (DOI)?
The data is distributed through a Microsoft OneDrive link, which includes a zip of the images themselves as well as spreadsheets for the labels.

When will the data set be distributed?
The data set was first released in May 2022 as a pre-print.

Will the data set 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.
Yes, the data set is distributed under a CC-BY-NC-SA license, restricting commercial usage.

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.
There are no fees or restrictions.

Do any export controls or other regulatory restrictions apply to the data set 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

Who will be supporting/hosting/maintaining the data set? The data set is hosted at the University of British Columbia using the institution's instance of Microsoft OneDrive.

How can the owner/curator/manager of the data set be contacted (e.g., email address)?
All questions and comments can be sent to Rohit Singla: rsingla@ece.ubc.ca

Is there an erratum? If so, please provide a link or other access point. All changes to the data set will be maintained under an erratum located at `https://github.com/rsingla92/kidneyUS/blob/main/README.md#errata`

Will the data set be updated (e.g., to correct labelling 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)?
All changes to the data set will be maintained on the website as well as through the registration list.

If the data set 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 data set continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.
They will continue to be supported with all information online unless otherwise communicated.

If others want to extend/augment/build on/contribute to the data set, 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
There currently is no such mechanism in place.

Any other comments?
n/a