Datasheet for FIRE-D: NASA-centric Remote Sensing of Wildfire

Yuzhou Chen
Temple University

Nicholas LaHaye
NASA's JPL at Caltech

Jae Won Choi
UT Dallas

Zhiwei Zhen
UT Dallas

Philip E. Davis
University of Utah

Huikyo Lee
NASA's JPL at Caltech

Manish Parashar
University of Utah

Yulia R. Gel
UT Dallas

A Datasheet

This Datasheets for Dataset follows the template from [2, 1].

A.1 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.

A: Satellite imagery is created for a variety of purposes, each serving different applications and academic research. A number of satellite-based instruments do not provide an operational per-pixel wildfire identification dataset. Recognizing the need, SIT-FUSE used these images to mark the regions where active fires are burning. By using these wildfire label data, machine learning models can learn to recognize patterns, features, and characteristics associated with wildfires, enabling them to autonomously detect and monitor such events in real-time or at scale.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A: Dr. Nicholas LaHaye, NASA's Jet Propulsion Lab (JPL) has pre-processed the satellite imagery from various sources and created the wildfire label data, as a part of the NASA AIST funding efforts on open benchmarks.

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.

A: This project has been supported by NASA Advanced Information Systems Technology (AIST) grants 21-AIST21_2-0020 and 21-AIST21_2-0059.

Any other comments?

A: None.

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

A.2 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.

A: The satellite imagery data include high-resolution images of the Earth's surface from space at different electromagnetic frequencies.

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

A: There are 157,469,891 samples labeled -1, 297,971,233 samples labeled 0, and 9,276 samples labeled 1.

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).

A: The temporal coverage of satellite observations is longer than that of those prepared for FIRE-D. Additionally, while satellite-based instruments observe large regions of the Earth, FIRE-D focuses solely on the contiguous United States (CONUS). The dataset over the CONUS does not represent the entire scope of observations but is still useful to support a variety of wildfire studies using ML models. This is because almost all ML models designed to predict active fires are built for specific geographcal regions.

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

A: High-resolution images.

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

A: We use JPL's Segmentation, Instance Tracking, and data Fusion Using multi-SEnsor imagery (SIT-FUSE), to generate a set of labels to mark actgive wildfires – previously validated on numerous pre-existing label sets for wildfires. For this task, each wildfire label set is generated at the input data's native resolution and contains 3 values: $-1 = no\_input\_data$, $0 = no\_fire$, and $1 = fire$.

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.

A: No information is missing.

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.

A: The location and time at which each image was independently taken are specified.

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.

A: In this paper, we conduct unsupervised learning tasks and we do not split data into training, development/validation, testing. The dataset can be used to evaluate sensitivity of model performance with respect to different splits.

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

A: Satellite image has noise caused by atmospheric interference, sensor artifacts, cloud cover, or other environmental conditions. This kind of noise in satellite image can affect the accuracy of the wildfire label, but this is beyond the scope of this work.

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 dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A: 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.

A: The dataset contains no confidential information.

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

A: The dataset contains no information that that, if viewed directly or indirectly, might be offensive, insulting, threatening, or cause anxiety.

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.

A: The dataset does not identify any subpopulations and their respective socio-demographic characteristics.

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.

A: It is not possible to identify individuals, either directly or indirectly from the dataset.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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.

A: The dataset does not contain any sensitive information.

Any other comments?

A: None.

A.3 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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A: Generating satellite images from raw observations involves pre-processing the data to correct errors and artifacts, and geo-referencing the data.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

A: The pre-processed and geo-referenced observations are reconstructed into an image format, forming a spatial grid where each pixel represents a location on Earth. Additional enhancement techniques may be applied to improve visual quality or highlight specific features. The procedure follows an algorithm theoretical basis document (ATBD) for each satellite mission, and the ATBD includes the validation of the procedures.

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

A: The image datasets are spatial and temporal subsets of the entire archive, with FIRE-D specifically focusing on CONUS.

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: Dr. Nicholas LaHaye, NASA's Jet Propulsion Lab (JPL) has collected the satellite imagery from various sources and created the wildfire label data, as part of his commitment to the NASA-funded AIST project on open benchmarks.

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.

This dataset contains 34 LandSat-8 scenes from September of 2022. Within the scenes there are 455,450,400 total samples.

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.

A: The ethical review process is not applicable to the collected data on Earth science processes.

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

A: Not applicable.

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.

A: Not applicable.

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.

A: Not applicable.

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).

A: Not applicable.

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.

A: The satellite image datasets are publicly available, so the dataset does not require its own impact assessment.

Any other comments?

A: None.

A.4 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 remaining questions in this section.

A: No pre-processing or cleaning is performed, but the data represent a certain time period over CONUS.

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.

A: The raw data are publicly available from:

`https://search.earthdata.nasa.gov/search?q=HLS%20Daily%20Global`, and `https://scihub.copernicus.eu/dhus/#/home`.

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

A: The software is publicly available here:

`https://github.com/nlahaye/SIT_FUSE/`.

Any other comments?

A: None.

A.5 Uses

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

A: Images from Earth observation satellites are used to generate scientific level 2 products. Level 2 products typically involve the processing and analysis of satellite image to derive specific parameters or measurements. These products are often more refined and standardized, catering to scientific research, analysis, and modeling in atmospheric sciences.

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.

A: GeoQuery is available at `https://github.com/186philip-davis/GeoQuery`.

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

A: The wildfire labels can be used as input for physics-based numerical models, which can simulate transport of wildfire-induced smoke plumes and their impact.

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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

A: Given that there has not been a major update in generating images from raw observations, the satellite imagery and wildfire labels can be used in their current form.

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

A: We do not anticipate any restrictions on the tasks that the dataset should not be used.

Any other comments?

A: None.

A.6 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.

A: The dataset is publicly available.

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

A: The wildfire labels dataset is available at `https://osf.io/v4uz9/?view_only=a6f58def39d344dba81952968ea234cd`

When will the dataset be distributed?

A: This dataset is already available.

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.

A: The access repository and dataset area both provided under the MIT license, and can be accessed without cost.

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.

A: We are not aware of any third-party restrictions or costs imposed.

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.

A: The dataset does not require export controls or other regulatory restrictions.

Any other comments?

A: None.

A.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

A: The dataset is hosted by the Open Science Foundation (`osf.io`). The dataset and access repo will be supported by Philip Davis at the University of Utah, as part of NASA AIST initiative on open benchmarks.

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

A: philip.davis@sci.utah.edu.

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

A: There is no erratum.

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 dataset consumers (e.g., mailing list, GitHub)?

A: Errors in the dataset will be corrected on an as-necessary basis. Updates will be communicated on the OSF repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the 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.

A: The dataset does not relate to people.

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 dataset consumers.

A: Older versions will not be hosted or maintained. Obsolescence will be communicated on the OSF repository, if necessary.

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 dataset consumers? If so, please provide a description.

A: Users may fork and modify the OSF and/or GeoQeury repository using common facilities for doing so.

Any other comments?

A: None.

References

[1] Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. *Advances in Neural Information Processing Systems*, 35:25979–25991, 2022.

[2] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.