

[pair-code.github.io/datacardsplaybook](http://pair-code.github.io/datacardsplaybook)  
Data Cards Template (Beta) | Updated November 2022

**ABRIDGED VERSION PRODUCED FOR I-GUIDE SUMMER SCHOOL August 2023**

|  |  |
| --- | --- |
| A black and white sign with a person in a circle  Description automatically generated | The [Data Cards Playbook ↗](http://pair-code.github.io/datacardsplaybook) by Google Research is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.  You are free to share and adapt this work under  the [appropriate license terms ↗](https://creativecommons.org/licenses/by-sa/4.0/deed.ast). |
| Team 6 dataset | The dataset utilized in this project contains temporal Harmonized Landsat-Sentinel imagery of diverse land cover and crop type classes across the Contiguous United States for the year 2022. The data card is right here <https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification>. The primary motivation behind this dataset was to use a model capable of generating masked datasets for specific classification, leveraging training datasets for model training and validation. The resultant dataset showcases prediction results that could be vital for urban planning, environmental studies, disaster management, and other relevant fields. The process's benefits include the ability to monitor land use changes, understand urban expansion, and contribute to sustainable development practices. This dataset is particularly suitable for applications that require a detailed understanding of land use patterns, where quality spatial information about built-up areas is essential. Its utilization promises to enhance the understanding and management of urban landscapes, catering to both scientific research and practical applications in various domains. |
| DATASET LINK | DATA CARD AUTHOR(S) |
| Dataset Link  <https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification/tree/main> | **Claire Simpson, Team 6:** (Contributor)  **Salar Jarhan, Team 6:** (Contributor)  **Yalin Yang, Team 6:** (Contributor)  **Jiyoung Lee, Team 6:** (Contributor)  **Yanhong Huang, Team 6:** (Contributor) |

|  |  |  |
| --- | --- | --- |
| Authorship | | |
| Dataset Owners | | |
| TEAM(S) | CONTACT DETAIL(S) | AUTHOR(S) |
| Team 6 | **Dataset Owner(s):** I-GUIDE  **Affiliation:** University of Illinois at Urbana-Champaign  **Contact:** help@cybergis.org  **Group Email:** help@cybergis.org  **Website:** https://iguide.illinois.edu/ | Claire Simpson, Ph.D. student, University of Colorado, Boulder, 2023  Salar Jarhan, Ph.D. student, University of Wyoming, 2023  Yalin Yang, Ph.D. student, University of Texas, Dellas, 2023  Jiyoung Lee, Ph.D. student, Lousina State University, 2023  Yanhong Huang, Ph.D. student, University of New Mexico, 2023 |

###### 

|  |  |  |
| --- | --- | --- |
| Dataset Overview | | |
| DATA SUBJECT(S) | DATASET SNAPSHOT | CONTENT DESCRIPTION |
| Non-Sensitive Data about peopleData about natural phenomena | |  |  | | --- | --- | | Size of Dataset |  | | Number of Instances | 4857 | | Number of Fields | 22 | | Each datapoint in the dataset consists of a satellite image from github, encompassing six specific bands: blue, green, red, NIR (Near-Infrared), SWIR1 (Short-Wave Infrared 1), and SWIR2 (Short-Wave Infrared 2). These bands are leveraged to calculate the NVBI (Normalized Built-up Index), a key metric instrumental in classifying land use. And also the results of the classification of the built-up and non built-up areas. |
| Sensitivity of Data | | |
| SENSITIVITY TYPE(S) | FIELD(S) WITH SENSITIVE DATA | SECURITY AND PRIVACY HANDLING |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Example of Data Points | | |
| PRIMARY DATA MODALITY |  | DATA FIELDS |
| Image Data |  |  |
| TYPICAL DATA POINT | | ATYPICAL DATA POINT |
|  | |  |

###### 

|  |  |  |
| --- | --- | --- |
| Motivations & Intentions | | |
| Motivations | | |
| PURPOSE(S) | RESEARCH AND PROBLEM SPACE(S) | MOTIVATING FACTOR(S) |
| Monitoring | The dataset serves as a foundation for developing a robust land classification model, addressing the critical need for efficient and accurate land use analysis. This has wide-reaching implications for urban planning, environmental conservation, and monitoring of human impact on the natural environment. | - Bringing landscape diversity to imagery training data for land use and land cover models.  - Insight into the distribution of built-up land supports city planning and sustainable development, ensuring optimal land use while preserving ecological balance.  - Detailed land classification can assist in predicting vulnerabilities to natural disasters, such as floods or earthquakes, and in developing more effective response strategies. |
| Intended Use | | |
| DATASET USE(S) | SUITABLE USE CASE(S) | UNSUITABLE USE CASE(S) |
| Safe for research use | **[Urban planning]**: For cities aiming to expand or restructure, this dataset can provide detailed insights into existing land use, helping planners design sustainable urban environments.  **[Agricultural Monitoring]**: Agricultural agencies may find the data useful to differentiate between built-up areas and arable land, helping them in crop planning and land management.  **[Disaster Risk Assessment]**: Agencies responsible for disaster management can utilize the dataset to analyze vulnerability to natural disasters like floods or landslides, guiding mitigation efforts. | **[Privacy Concerns]**: If the dataset is used to monitor private properties or sensitive locations without proper legal permissions or public disclosure, it could violate privacy rights.  **[Bias and Discrimination]**: If used improperly or without consideration of local context and community needs, the application of this data might favor certain groups over others, leading to biased development policies or resource allocations.  **[Military or Malicious Usage]**: The detailed land classification might be misused by military or malicious actors to identify strategic locations for operations that could harm individuals or national security. |

|  |  |  |
| --- | --- | --- |
| Provenance | | |
| Collection | | |
| METHOD(S) USED | METHODOLOGY DETAIL(S) | SOURCE DESCRIPTION(S) |
| Taken from other existing datasets | **Source:** The Harmonized Landsat Sentinel-2 (HLS) project is a NASA initiative that aims to provide consistent and ready-to-use data by combining observations from the U.S. Landsat and the European Union Sentinel-2 satellites.  **Platform:** [Jupyter notebook],  Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text.  **Is this source considered sensitive or high-risk?** [ No]  **Dates of Collection:** [01/2022 – 12/2022]  **Primary modality of collected data:**  Image Data  **Update Frequency for collected data:**  5-day | **[Github]:** GitHub is a web-based platform that provides hosting for software development and a collaborative work environment. It is one of the largest code repositories in the world, allowing developers to work on projects collaboratively. GitHub offers features like version control using Git, collaborative code review, issue tracking, and more. Open-source projects and private repositories can be hosted on the platform, and it has become a central hub for developers, researchers, and organizations to share code, collaborate on development, and build software together. GitHub plays a vital role in the open-source community, facilitating collaboration and innovation across various domains, including machine learning and data science.  **Additional Notes:** <Add here> |
| COLLECTION CADENCE | DATA INTEGRATION | DATA PROCESSING |
| Static (Data was collected once from single or multiple sources.) | **Included Fields**  (Data fields that were collected and are included in the dataset.)   |  |  | | --- | --- | | **Field Name** | **Description** | | B8A | NIR band | | B11 | SWIR1 band |   **Excluded Fields**  (Data fields that were collected but are excluded from the dataset.)   |  |  | | --- | --- | | **Field Name** | **Description** | | B02 | Blue band | | B03 | Green band | | B04 | Red band | | B12 | SWIR2 band | | **Description:** This dataset contains temporal Harmonized Landsat-Sentinel imagery of diverse land cover and crop type classes across the Contiguous United States for the year 2022. The target labels are derived from USDA's Crop Data Layer (CDL). It's primary purpose is for training segmentation geospatial machine learning models.  **Methods employed:** In Jupyter notebook, we used the code “wget -r <https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification/resolve/main/training_chips.tgz>” |

|  |  |  |
| --- | --- | --- |
| Human and Other Sensitive Attributes | | |
| SENSITIVE HUMAN ATTRIBUTE(S) | INTENTIONALITY | METHODOLOGY DETAIL(S) |
|  |  |  |
| DISTRIBUTION(S) | | |
| Human Attribute | | |
|  | KNOWN CORRELATIONS |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Extended Use | | |
| Use with Other Data | | |
| SAFETY LEVEL | KNOWN SAFE DATASET(S) OR DATA TYPE(S) | BEST PRACTICES |
|  |  |  |
|  | KNOWN UNSAFE DATASET(S) OR DATA TYPE(S) | LIMITATION(S) AND RECOMMENDATION(S) |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Transformations | | |
| Synopsis | | |
| TRANSFORMATION(S) APPLIED | FIELD(S) TRANSFORMED | LIBRARY(IES) AND METHOD(S) USED |
| Data AggregationDimensionality Reduction | **Data aggregation and Dimensionality Reduction**   |  |  | | --- | --- | | **Field Name** | **Source & Target** | | Band 04 | DN value: NDBI | | Band 05 | DN value: NDBI | | **Data aggregation and Dimensionality Reduction**  **Method:**  Selecting the Bands: From the multispectral data, select the NIR band and the SWIR band.  Calculating NDBI: The NDBI is calculated using the following formula:  NDBI=(SWIR−NIR)/(SWIR+NIR)  ​SWIR: Reflectance value of the SWIR band.  NIR: Reflectance value of the NIR band.  Interpretation: The NDBI values typically range from -1 to +1. Positive values often correspond to built-up areas, while negative values usually represent non-built-up areas.  Application: The NDBI can be used for mapping and analyzing urban expansion, land use changes, and spatial planning.  **Platforms, tools, or libraries:**  [I-GUIDE]: <NSF Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) enables transformative discovery and innovation for tackling fundamental scientific and societal challenges that are at the cusp of achieving significant breakthroughs by harnessing the vast, diverse, and ever-growing corpus of geospatial data.  [Jupyter notebook]: Jupyter Notebook is an open-source web application that allows you to create and share documents containing live code, equations, visualizations, and narrative text.  [rasterio]: Rasterio is a Python library that simplifies reading and writing raster datasets, making it easier to work with geospatial raster data. It's a highly valuable tool for anyone dealing with geographical information systems (GIS), remote sensing, or any field that requires handling raster data.  **Transformation Results:**  The calculated NDBI values provide a map that differentiates between built-up areas (urban regions) and non-built-up areas (such as vegetation and water). |
| Breakdown of Transformations | | |
| CLEANING MISSING VALUE(S) | METHOD(S) USED | COMPARATIVE SUMMARY |
|  |  |  |
| CLEANING MISMATCHED VALUE(S) | METHOD(S) USED | COMPARATIVE SUMMARY |
|  |  |  |
| ANOMALIES | METHOD(S) USED | COMPARATIVE SUMMARY |
|  |  |  |
| DIMENSIONALITY REDUCTION | METHOD(S) USED | COMPARATIVE SUMMARY |
| How many original features were collected and how many dimensions were reduced? | What methods were used to reduce the dimensionality of the data? | Why were features reduced using this method (over others)? Provide comparative charts showing before and after dimensionality reduction processes. |
| In this project, six original spectral bands were collected from the satellite images, encompassing blue, green, red, NIR (Near Infrared), SWIR1 (Short Wave Infrared 1), and SWIR2 (Short Wave Infrared 2). However, only two of these bands, specifically the NIR and SWIR bands, were used to calculate the Normalized Difference Built-up Index (NDBI).  **Field Name:** Count or Description  **Field Name:** Count or Description  **Field Name:** Count or Description | <Summarize here. Include links where necessary.>  **Platforms, tools, or libraries:**  **[Platform, tool or library]:** <Write description here.>  **[Platform, tool or library]:** <Write description here.>  **[Platform, tool or library]:** <Write description here.> | <Summarize here. Include links, tables, visualizations where available>   |  |  | | --- | --- | | **Field Name** | **Diff** | | <Field Name> | <Before: After> | | <Field Name> | <Before: After> | | … | … |   **Above:** <Provide a caption for the above table or visualization.>  **Additional Notes:** <Add here> |
| JOINING INPUT SOURCES | METHOD(S) USED | COMPARATIVE SUMMARY |
|  |  |  |