Datasheet for Solar Flare Prediction

Dataset Composition

**What are the instances?**

Each instance is a collection of measurements taken at a particular point in time.

**Are relationships between instances made explicit in the data**

N/A

**How many instances of each type are there?**

**What data does each instance consist of?** There are 31 features representing a timestamp and various measurements taken at that time. There is a target which indicates whether a flare resulted.

**Is everything included or does the data rely on external resources?**

No external resource required.

**Are there recommended data splits or evaluation measures?**

No.

**What experiments were initially run on this dataset?**

Unknown

**Any other comments?**

Data Collection Process

**How was the data collected?**

Using the Sunpy library, data was collected from the Joint Science Operations Center (JSOC) and Space Weather Prediction Center (SWPC).

**Who was involved in the data collection process?** Alciomar Hollanda, Ana Estela Antunes da Silva, Tiago Cinto

**Over what time-frame was the data collected?**

The data was collected from May 2010 to December 2019.

**How was the data associated with each instance ac- quired?**

The data was acquired from the HMI instrument aboard the [Solar Dynamics Observatory](https://en.wikipedia.org/wiki/Solar_Dynamics_Observatory). It is assumed that the authors perform some preprocessing on the data prior to release.

**Does the dataset contain all possible instances?**

It is a sample.

**If the dataset is a sample, then what is the population?**

Unknown

**Is there information missing from the dataset and why?**

There are 15 missing values for MEANSHR (Mean photospheric magnetic shear angle).

**Are there any known errors, sources of noise, or redundancies in the data?**

Unknown

**Any other comments?**

Motivation

**Why was the dataset created?**

To enable the training of machine learn models to predict solar flares.

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

N/A

**Has the dataset been used for any tasks already?** Unknown

**Who funded the creation of the dataset?**

Uncertain. Potentially the University of Campinas

**Any other comments?**

Legal & Ethical Considerations

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection?**

N/A

**If it relates to other ethically protected subjects, have appropriate obligations been met?**

N/A

**If it relates to people, were there any ethical review ap- plications/reviews/approvals?**

N/A

**If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?**

N/A

**If it relates to people, could this dataset expose people to harm or legal action?**

N/A

**If it relates to people, does it unfairly advantage or dis- advantage a particular social group?**

N/A

**If it relates to people, were they provided with privacy guarantees?**

N/A

**Does the dataset comply with the EU General Data Protection Regulation (GDPR)?**

N/A

**Does the dataset contain information that might be considered sensitive or confidential?**

No

**Does the dataset contain information that might be considered inappropriate or offensive?**

No

**Any other comments?**

Data Preprocessing

**What preprocessing/cleaning was done?**

Unknown.

**Was the “raw” data saved in addition to the preprocessed/cleaned data?**

It is assumed that the raw data is still available at Joint Science Operations Center (JSOC) and Space Weather Prediction Center (SWPC).

**Is the preprocessing software available?**

Unknown.

**Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?**

**Any other comments?**

Dataset Distribution

**How is the dataset distributed?**

[Website](https://zenodo.org/records/4603412)

**When will the dataset be released/first distributed?**

First published March 13, 2021

**What license (if any) is it distributed under?**

Public

**Are there any fees or access/export restrictions?**

No

**Any other comments?**

Dataset Maintenance

**Who is supporting/hosting/maintaining the dataset?**

Unknown

**Will the dataset be updated?**

Unknown

**If the dataset becomes obsolete how will this be communicated?**

Unknown

**Is there a repository to link to any/all papers/systems that use this dataset?**

Unknown

**If others want to extend/augment/build on this dataset, is there a mechanism for them to do so?**

Unknown

**Any other comments?**

Appendix 1 – Definition of fields

ABSNJZH - Absolute value of the net magnetic flux in the vertical component.

AREA\_ACR - The area of the active region in millionths of a solar hemisphere.

EPSX, EPSY, EPSZ - Electric current helicity in the east-west, north-south, and vertical directions.

Latitude - The angular distance on the Sun's surface measured north or south of the solar equator.

Longitude - The angular distance on the Sun's surface measured east or west of the central meridian

NOAA\_AR - This is the unique identifier assigned by the National Oceanic and Atmospheric Administration (NOAA) to a specific active region on the Sun.

MEANALP - Mean inclination angle of the magnetic field to the local vertical.

MEANGAM - Mean inclination angle of the magnetic field to the local vertical.

MEANGBH, MEANGBT, MEANGBZ - Mean magnetic field strength in the total(T), vertical(Z), and horizontal(H) components.

MEANJZD - Mean current density in the north-south direction.

MEANJZH - Mean current helicity, a measure of the twist and shear in the magnetic field.

MEANPOT - Mean magnetic potential field strength in the line-of-sight (vertical) direction.

MEANSHR - Mean photospheric magnetic shear angle.

QUALITY - Denotes the quality of the observation or measurement for that active region, indicating the level of confidence in the data.

R\_VALUE - Measures the complexity of the magnetic field configuration.

SAVNCPP - Sum of the absolute value of the net magnetic flux in the line-of-sight direction for all sunspots within the active region.

SHRGT45 - Fraction of pixels with shear angles greater than 45 degrees.

TOTFX, TOTFY, TOTFZ - Total flux in the east-west direction(X), north-south direction(Y), line-of-sight (vertical) direction(Z).

TOTBSQ, TOTPOT - Total unsigned magnetic flux and total magnetic potential field strength. These give additional information about the magnetic configuration.

TOTUSJH, TOTUSJZ - Total unsigned magnetic flux in the horizontal(H) and vertical(Z) components. These provide an overall measure of the magnetic field strength in the active region.

USFLUX - Unsigned magnetic flux.