SOLAR ACTIVE REGION MAGNETOGRAM IMAGE DATASET FOR STUDIES OF SPACE WEATHER

Laura E. Boucheron
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, NM 88001, USA
lboucher@nmsu.edu

Ty Vincent
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, NM 88001, USA

Jeremy A. Grajeda
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, NM 88001, USA
jgra@nmsu.edu

Ellery Wuest
Klipsch School of Electrical and Computer Engineering
New Mexico State University
Las Cruces, NM 88001, USA
ellerywu@nmsu.edu

ABSTRACT

In this dataset we provide a comprehensive collection of magnetograms (images quantifying the strength of the magnetic field) from the National Aeronautics and Space Administration's (NASA's) Solar Dynamics Observatory (SDO). The dataset incorporates data from three sources and provides SDO Helioseismic and Magnetic Imager (HMI) magnetograms of solar active regions (regions of large magnetic flux, generally the source of eruptive events) as well as labels of corresponding flaring activity. This dataset will be useful for image analysis or solar physics research related to magnetic structure, its evolution over time, and its relation to solar flares. The dataset will be of interest to those researchers investigating automated solar flare prediction methods, including supervised and unsupervised machine learning (classical and deep), binary and multi-class classification, and regression. This dataset is a minimally processed, user configurable dataset of consistently sized images of solar active regions that can serve as a benchmark dataset for solar flare prediction research.

1 Background & Summary

In this dataset, we provide a comprehensive collection of magnetograms (images quantifying the strength of the magnetic field) from the National Aeronautics Space Administration's (NASA's) Solar Dynamics Observatory (SDO). SDO was launched on 11 February 2010 as the first mission in support of the Living With a Star (LWS) program which seeks to understand solar variability and the effects of space weather at Earth and throughout the Solar System [1]. Specific goals of SDO in line with this dataset are to better understand the magnetic structure of the Sun and understand and predict how that magnetic structure initiates space weather events such as flares [1]. Three experiments are included on

SDO: the Atmospheric Imaging Assembly (AIA) [2], the EUV Variability Experiment (EVE) [3], and the Helioseismic and Magnetic Imager (HMI) [4]. In this paper, we focus on line-of-sight magnetogram images from HMI.

The dataset presented in this paper provides a comprehensive set of HMI magnetograms of solar active regions (regions of large magnetic flux, generally the source of eruptive events) as well as labels of corresponding flaring activity. This dataset will be useful for image analysis or solar physics research related to magnetic structure, its evolution over time, and its relation to solar flares (a sudden and large emission of radiation). It is expected that the main audience for this dataset are those researchers investigating automated solar flare prediction methods, including supervised and unsupervised machine learning (classical and deep), binary and multi-class classification, and regression. While SDO provides an incredibly rich dataset that can be an excellent source for image processing and machine learning researchers, there are several characteristics of the data that motivated our creation of this specific dataset. First, and overarching, was the desire to provide a minimally processed, user configurable dataset that can serve as a benchmark dataset for solar flare prediction research. Second was the desire to focus analysis on solely active regions and to reduce the amount of time needed to interact with existing interfaces to download such data. Third was the desire that images of those active regions be consistently sized images rather than varying across active regions and/or across time. Fourth was the necessity of integrating a separate dataset in order to develop labels related to flare activity.

In this dataset, we address the aforementioned characteristics as follows. First, we provide a comprehensive set of magnetogram images from all National Oceanic and Atmospheric Administration (NOAA) active regions (ARs) from May 2010 through December 2018. Along with this set of images, we provide a means to configure basic parameters of the dataset, including the size of flares to consider, the time window over which to consider flare prediction, the latitudes and longitudes of active regions to include, and whether to include images with Not-a-Number (NaN) pixel values. Second, we integrate two sources of data in order to retrieve data only associated with ARs and provide a means to automate the download of those AR magnetogram images. Third, we provide consistently sized (600×600 pixel) images, which can be an important assumption in batch processing of images, particularly for some common deep learning methods, e.g., convolutional neural networks. Fourth, we integrate a third source of data in order to provide labels related to flaring activity.

2 Methods

2.1 Dataset Overview

This dataset incorporates data from three main sources. First, in order to focus the image collection on ARs, we used the NOAA Space Weather Prediction Center (SWPC) Solar Region Summaries (SRS) (`ftp://ftp.swpc.noaa.gov/pub/warehouse/`) and parsed those text data to extract the date an AR appeared on disk and the number of days it was visible on disk. Additionally, the SRS provide latitude and longitude of ARs which we use to postprocess the dataset. Second, we download magnetogram images from SDO/HMI using the Joint Science Operations Center (JSOC) interface (`http://jsoc.stanford.edu/ajax/lookdata.html`) at a cadence of 720 seconds, centered at the NOAA AR centroid (tracked according to the Carrington rate), and with a spatial extent of 600×600 pixels. This image size was chosen to correspond to approximately 300 arcseconds $\times$ 300 arcseconds ($300'' \times 300''$) commensurate with previous work on solar flare prediction, e.g., [5, 6], and to be large enough to encompass the typical range of AR sizes [7]. Third, we used the SWPC Event Reports (ER) (`ftp://ftp.swpc.noaa.gov/pub/warehouse/`) to extract the AR number, peak flare time, and flare size in order to provide labels for those researchers investigating a supervised classification or regression problem. Figure 1 summarizes the data flow used to create this dataset.

In total, we downloaded images corresponding to 1,655 NOAA ARs which appeared with sunspot structure on the Sun from 01 May 2010 through 31 December 2018, a total of 1,372,004 HMI images from NOAA ARs 11064 through 12731. We only include those ARs which appeared for the totality of their lifetime within the time range 01 May 2010 through 31 December 2018; thus ARs which were already present on the Sun prior to 01 May 2010 or continued their presence on the Sun after 31 December 2018 are not included in this dataset. NOAA ARs 11160, 11171, 12623, and 12705 never developed sunspots and thus contribute no images to this dataset. Additionally, NOAA ARs 11190, 11493, 11494, 11496, 11501, 11503, 12472, 12473, and 12570 are not included in this dataset since they appeared during times when the SDO satellite was missing fine guidance [8] and thus the location of the ARs could not be accurately tracked. (More specifically, a reference time (`http://jsoc.stanford.edu/doxygen_html/im_patch_8c-source.html`) is specified for the AR corresponding to the time that AR will be at disk center (`http://jsoc.stanford.edu/doxygen_html/libs_2astro_2heliographic__coords_8c-source.html`) and no data records are returned if there are no valid data within a four hour window of that reference time.) The entire image set (i.e., the 1,372,004 `.fits` images) comprises 537 GB. We also provide a pre-configured AR dataset of `.fits` images and corresponding flare labels, described below, which comprises 375 GB and a reduced size (spatially and bit-depth) dataset of `.png` images and corresponding flare labels, also described below, which comprises 15 GB.
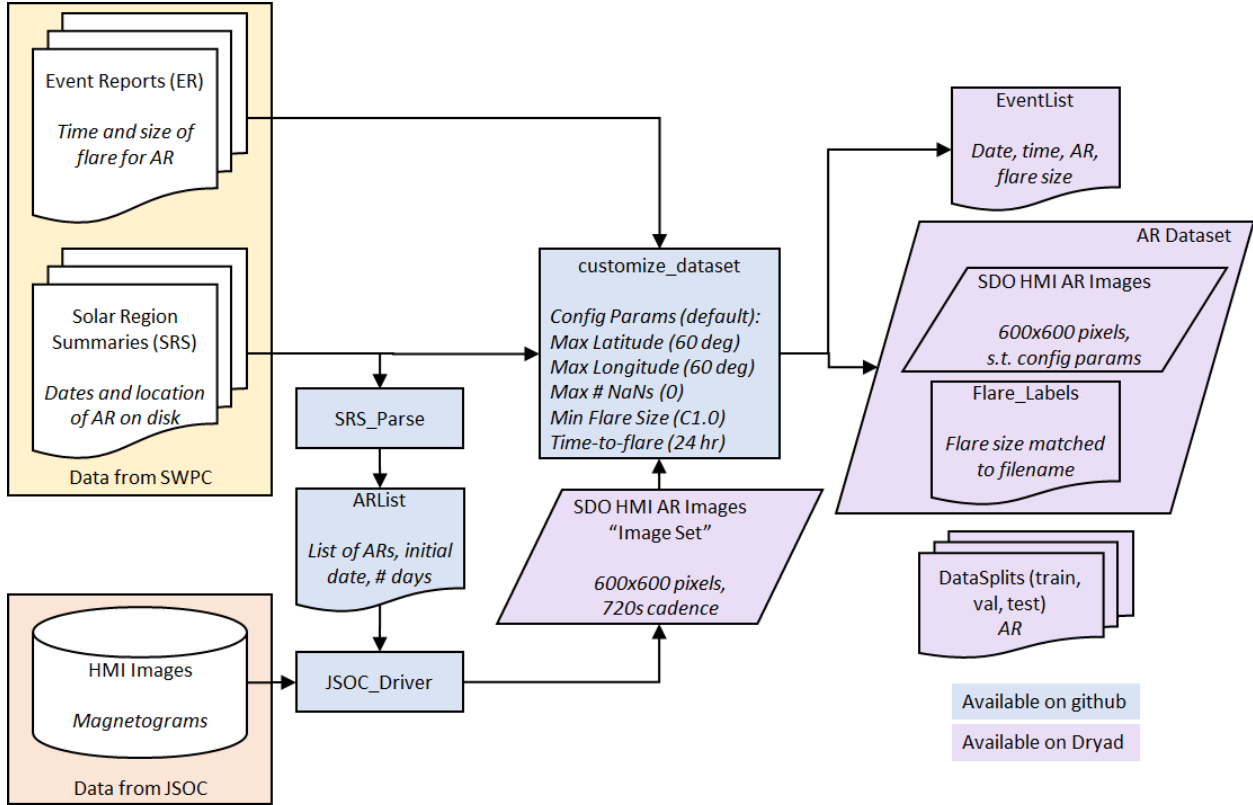
Figure 1: Flowchart of dataset creation. Space Weather Prediction Center (SWPC) Solar Region Summaries (SRS) are used to determine the dates for which a National Oceanic and Atmospheric Administration (NOAA) Active Region (AR) is visible on disk. Solar Dynamics Observatory (SDO) Helioseismic and Magnetic Imager (HMI) magnetogram images of ARs are downloaded via the Joint Science Operations Center (JSOC) web interface. SWPC Event Reports (ER) are used to specify the time and size of solar flares associated with a given NOAA AR.

2.2 Parsing the Solar Region Summaries for Active Regions

We used the NOAA SWPC SRS (`ftp://ftp.swpc.noaa.gov/pub/warehouse/`) to determine the dates a NOAA AR is visible on disk. The SRS are downloaded as one `.txt` file per day. We used `Part I` data in the SRS which detail those active regions with associated sunspot structures [9]. For each NOAA AR appearing in SRS `Part I`, we store the NOAA AR number, the date the AR first appears in the SRS, and accumulate the total number of days the same AR appears in the SRS. We store these data in a comma separated text file `ARList.txt` where each line is of the format `NNNN,YYYYMMDD,X`, where `NNNN` is the four digit NOAA AR number, `YYYYMMDD` is the initial date of appearance, and `X` is an integer number of days. The `AR_List.txt` file used to download the image set described here is provided as part of the GitHub repository at [10].

2.3 Downloading the Magnetograms for Active Regions

The text file `AR_List.txt` as described above is used to specify an appropriate date range to download the HMI magnetograms centered on a given AR. We request HMI magnetograms beginning at time 00:00:00 on the first day the AR appeared through 00:00:00 on the first day the AR disappeared. While there are modules to access SDO data for python (e.g., sunpy [11]) and IDL (e.g., SolarSoft [12]) without navigating the JSOC webpage, the ability to extract and track a cutout around a NOAA AR does not appear to be accessible through any means other than the website. In order to automate this process to download the 1,655 ARs, we wrote a python script to interact with the webpage using the selenium package [13] and geckodriver [14] for Mozilla's firefox web browser. We provide this code as part of the GitHub repository [10], but note that the code will break if any of the underlying html code on the JSOC website changes.

Since the JSOC driver code is fragile, we describe in detail the process of interacting with the JSOC Data Export webpage to download a single AR of data here. Readers who are interested in using the curated datasets [15, 16, 17] described in this paper can skip to the next subsection. Readers who are interested in downloading a custom dataset from the JSCO Data Export webpage may be interested in the process described here. This process assumes that the SWPC SRS have been parsed as in the previous section to determine the beginning date and number of days the AR is on disk.

- Navigate to the JSOC Data Export tool: `http://jsoc.stanford.edu/ajax/exportdata.html`

- In the `RecordSet` field enter the data locator in the form

  <div align="center">

  `hmi.M_720s[date1_time1_TAI-date2_time2_TAI][?quality>=0?]`

  </div>

  where dates and times are in the format `YYYY.MM.DD_HH:MM:SS`, `TAI` is the designation for international atomic time used by SDO, and the `quality` keyword specifies a search only for observables that were created. Press enter and the `Record Count` field will change to the total number of images spanned by the requested time period. There should be approximately 120 images per day requested.

- Using the `Method` dropdown menu, select `url-tar`.

- Check the `Enable Processing` checkbox which will result in the appearance of several additional check boxes.

- Check the `im_patch` checkbox which will result in the appearance of an `Image Patch Extract` box.

- In the `Image Patch Extract` box:

  - Ensure `Tracking` is checked in the `options` row.
  - Specify the `NOAA AR number` in the `options` row as a four or five digit number. Press enter and the `T_REF`, `X`, and `Y` fields will populate with reference time and location information for the AR. If the four digit truncated NOAA AR number is entered, the field automatically changes to the corresponding five digit number.
  - Verify `T_START` and `T_STOP` match the dates given in the `RecordSet` field.
  - Verify `Cadence` matches the cadence set in the `RecordSet` field.
  - Verify `BoxUnits` is set to `pixels`.
  - Set `Width` and `Height` to 600 each.
  - Click the `Check Params` button which will change the adjacent text field from `Not Ready` to `OK to submit`

- Verify `Protocol` is set to `FITS`.

- Enter the user's email (to which notification will be sent when the data is ready to be downloaded) in `Notify` field and user's name in `Requestor` field. The user's email must match a registered user (see also next bullet).

- Click `Check params for export` and the `Not Ready To Submit` button will change to a `Submit Export Request` button. If the email entered in the `Notify` field is not registered, a message will appear specifying that the user should respond to an email from JSOC within 15 minutes to register their email. An email will be sent from jsoc@sun.Stanford.EDU with subject "CONFIRM EXPORT ADDRESS" with further instructions. In short, a simple response to that email will register the user after which the user should receive a second email with subject "EXPORT ADDRESS REGISTERED." After this initial registration process, the user will need to click on the `Check params for export` button again. This registration process will need to be completed only once per user.

- Click `Submit Export Request` at which point the `RequestID` field will be populated with a string used to identify the data request. There may be few second delay before the `RequestID` field will populate.

- At the bottom of the page in the `JSOC Data Export Status and Retrieval` section, verify `RequestID` matches the above given `RequestID`.

- Periodically click `Submit Status Request` until the `Status` field becomes `Data Ready`. The `Status` may say `Bad Request Status` for the first few clicks of `Submit Status Request`; continue to click the same button until a request time is displayed in the `Status` field.

- When the `Status` field becomes `Data Ready`, click on the link provided in the `TarFile Location` field to download the requested data.
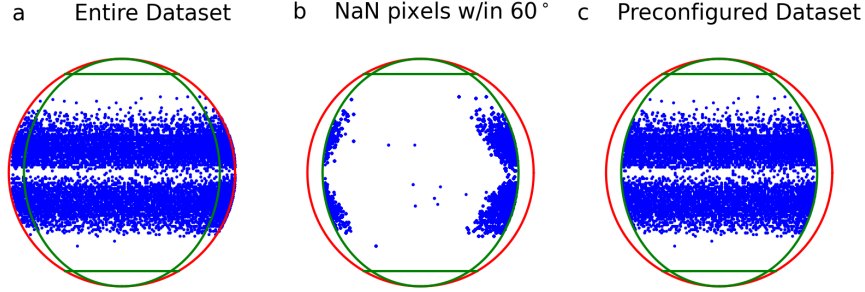
Figure 2: Latitude and longitude of AR images. The red circle denotes the solar radius and the green lines denote $\pm 60°$ latitude and longitude. The blue dots denote the centroids of the ARs included in the respective datasets. a: Latitude and longitude of files for entire dataset (image set). b: Latitude and longitude of files within $\pm 60°$ and ≥ 1 NaN pixels. c: Latitude and longitude of files for preconfigured AR dataset.

2.4 Parsing the Event Reports for Active Regions

Using the SWPC Event Reports (ER) [18] we parsed the text data for `XRA` events in the `Type` column (corresponding to x-ray events detected by the Geospatial Operational Environmental Satellite (GOES) spacecraft) with an associated number in the `REG#` column (corresponding to a NOAA AR number). For those x-ray events associated with a NOAA AR, we additionally parsed the ER for the peak flare time (`Max` column), and flare size (`Particulars` column). We store these data in a comma separated text file `Event_List.txt` where each line is of the format `YYYY MM DD,HHMM,NNNN,KX.X` where `YYYY MM DD` is the date, `HHMM` is the time, `NNNN` is the four-digit NOAA AR number, and `KX.X` is the GOES flare class (e.g., `C1.0` or `X10.1`) [19]. The `Event_List.txt` file for this dataset is provided as part of the image set at [17].

2.5 Customizing the Dataset

In this section we provide details on the postprocessing of the dataset according to AR location and flaring behavior. We provide a preconfigured dataset consisting of AR magnetograms within $\pm 60°$ latitude and longitude, containing zero NaN pixels, and labeled according to flaring behavior within 24 hours and at a flare size greater than C1.0. As described above, we download magnetogram images for NOAA ARs for the duration of their appearance on the solar disk; hereafter, we refer to this as the "image set" to distinguish it from the "AR dataset." The preconfigured AR dataset (described below) is available at [15] and a reduced resolution preconfigured AR dataset (described below) is available at [16]. The image set can be acquired by combining the preconfigured AR dataset [15] and the extra images dataset [17] which contains those images removed in the preconfiguration process.

2.5.1 Filtering Data by Latitude, Longitude and Not-a-Number (NaN) Pixels

Figure 2a shows a scatter plot of the latitude and longitude of the AR centroids for the image set. Some of these images, however, are near the edge of the solar disk and parts of the image capture data from off the solar disk (see Figure 3a). These disk-edge images may contain nonsensical magnetic measurements or NaN values. Furthermore, since the HMI magnetograms are line-of-sight (LOS), edge-of-disk images are affected by larger projection effects. These projection effects depend not only on the viewing angle but also on the specific geometry of the magnetic field, with deviations from radial in regions of stronger magnetic field introducing larger projection errors [20]. In this dataset, we do not implement any correction for projection effects, e.g., those in [20], but do provide a means for the user to configure a dataset by restricting the resulting images to reside within latitude and longitude bounds to limit the errors introduced by projection effects. We further note that the user could apply additional preprocessing methods to any of the image set images.

We use the SRS to determine the latitude and longitude for an AR on a given date, noting that the latitude and longitude are provided in the SRS at a daily cadence. Thus, we may exclude some images near the east limb that are just outside of the longitude threshold and rotate into a valid range throughout the day. Similarly, we may include some images near the west limb that are just inside the longitude threshold and rotate out of the valid range throughout the day. Using the daily latitude and longitude provided in the SRS files, we include in the preconfigured AR dataset all images with an AR centroid within $\pm 60°$ latitude and longitude (similar to those data in [5, 6]). A total of 313,601 files, comprising 22.9% of the entire dataset, are excluded from the preconfigured AR dataset based on latitude and longitude; a total of 85 ARs are excluded entirely based on these criteria.
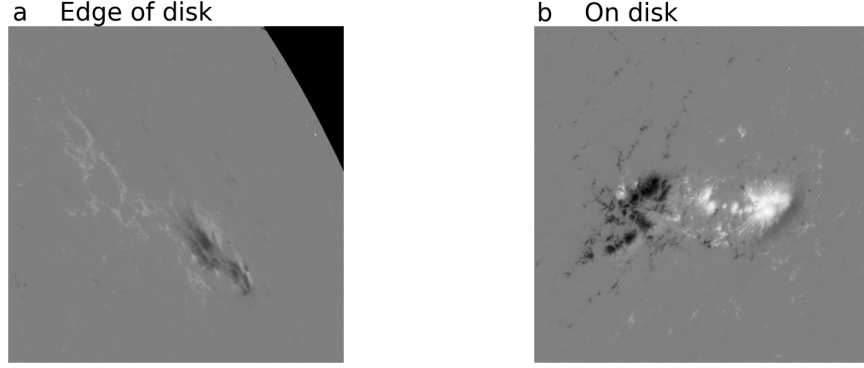
Figure 3: Examples of 600×600 pixel magnetogram images, including a disk-edge magnetogram and an on-disk magnetogram. a: Disk-edge magnetogram. NOAA AR 1169, 2011 March 15, 12:00:00. b: On-disk magnetogram. NOAA AR 2396, 2015 August 11, 00:00:00.

Due to the constant 600×600 pixel window of the images, ARs further from the equator may still contain off-disk data and we additionally exclude any image containing any NaN values, an additional 108,356 files and 7.9% of the entire dataset. The majority of these images with NaN values contain a small portion of the disk edge, but there are some images with spurious NaN values from various latitudes and longitudes. Figure 2b shows a scatter plot of those ARs within $\pm 60°$ latitude and longitude which contained at least one NaN pixel. We note that the majority of these images are near the disk edge, with a higher number of these images clustered near the west limb as compared to the east limb. This is consistent with the expectation that active regions on the west limb will be rotating closer to the disk edge throughout the day.

In total, between the latitude/longitude filtering and the NaN filtering, we exclude 421,957 images, comprising 30.8% of the entire dataset, from the preconfigured dataset. This results in a preconfigured dataset consisting of 950,047 on-disk HMI images (see Figure 3b) within a range of latitudes and longitudes (see Figure 2c). We provide the 950,047 images as part of the preconfigured AR dataset [15] and the reduced resolution dataset [16].

2.5.2 Assigning Flare Labels to Images

In order to use the dataset for supervised classification or regression, each image in the AR dataset needs a corresponding label specifying whether that image is associated with a flare. We provide a label indicating the flare size (as a string of GOES class, e.g., `'C1.0'`) for images associated with flares or `'0'` for images associated with non-flaring behavior. The user can configure the minimum flare size as well as the temporal flare prediction window; any images within the prediction window leading up to a flare are associated with that flare. For those ARs that flare multiple times within the flare prediction window, images are assigned a class associated with the largest size flare, consistent with [6].

Figure 4a shows a plot of the number of C-, M-, and X-class flares during the timespan of this dataset, while Figures 4b and 4c show counts of images associated with flaring behavior for a 24 hour flare prediction window for the entire dataset. We notice very similar trends in the count of flare events (Figure 4a) and the count of files associated with a flare (Figure 4b). This indicates that the entire dataset has well-sampled the flaring behavior of the Sun over this time period.

In order to assign labels to the AR dataset images, we loop over each event in `Event_List.txt` and assign a label of the GOES flare size for all images of the AR within 24 hours of the peak flare time for any flare size that satisfies the user-specified minimum flare size. After assigning flaring images for all events in `Event_List.txt`, all remaining images are labeled `'0'` to denote non-flaring images. The flare labels are stored in a file `KX.X_Hhr_Labels.txt` file where `KX.X` is the user-specified minimum flare size, e.g., `C1.0`, and `H` is the user-specified prediction window in hours, e.g., `24`. Each line in the flare labels file is of the form `filename,label` where `filename` is the base filename and `label` is the label (flare size for flaring and `'0'` for non-flaring).

For the preconfigured AR dataset, we specify a 24 hour prediction window and a minimum flare size of C1.0. We provide the `C1.0_24hr_Labels.txt` file as part of the preconfigured AR dataset [15] and the `C1.0_24hr_png_Labels.txt` file as part of the preconfigured reduced resolution AR datset [16], both of which contain 190,582 flaring images and 759,465 non-flaring images. Figures 4d and 4e show plots of images associated with flaring behavior for the preconfigured AR dataset. We notice very similar trends in the count of flare events for the entire dataset (Figures 4b and 4c) and in the preconfigured AR dataset (Figures 4d and 4e). This indicates that the configuration of the
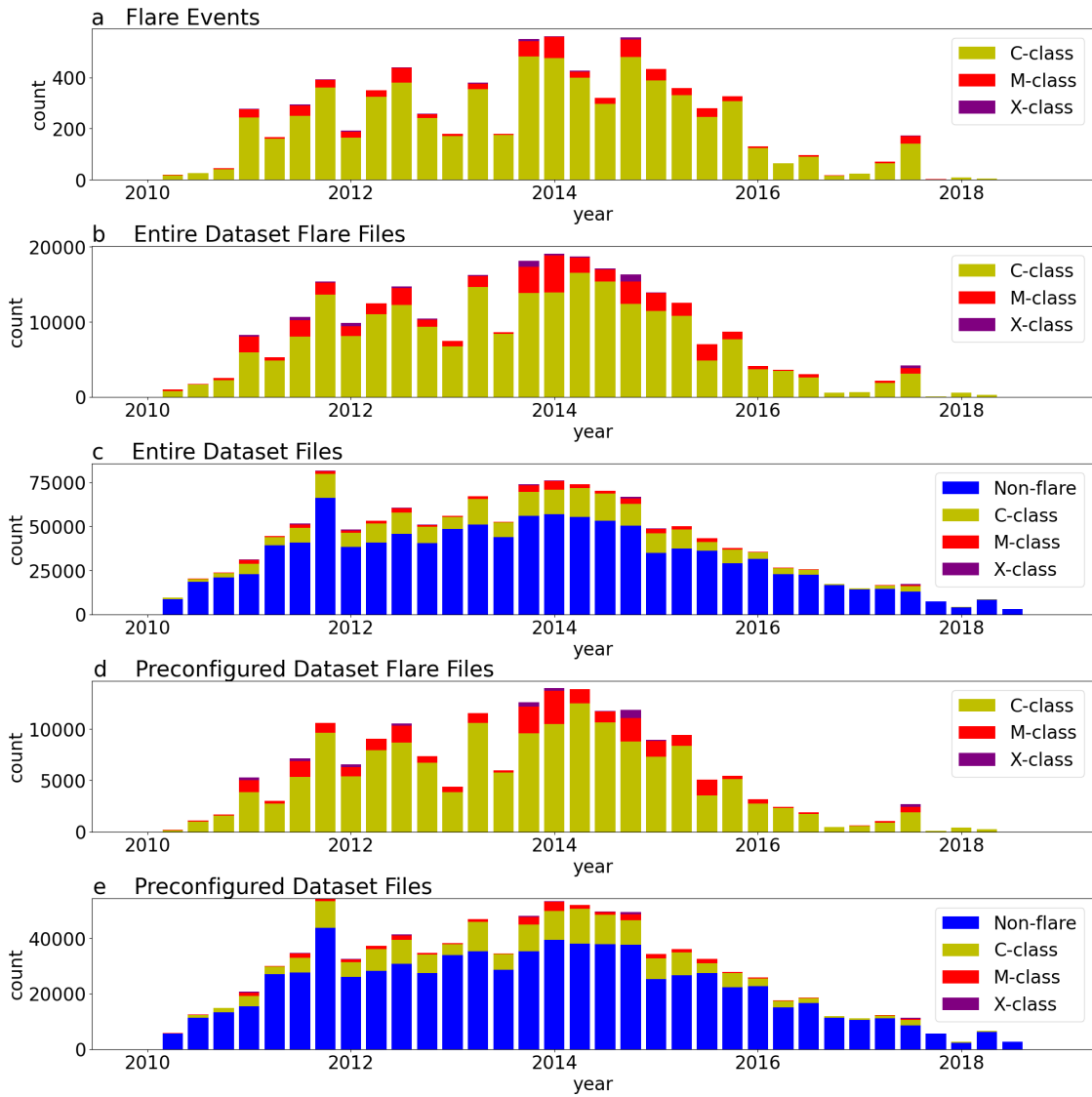
Figure 4: Count of events or files for different flaring behavior versus quarter. a: Count of flare events in the entire dataset (image set). b: Flare file count for the entire dataset (image set). c: Flare and non-flare file count for the entire dataset (image set). d: Flare file count for the preconfigured dataset. e: Flare and non-flare file count for the preconfigured dataset.

preconfigured AR dataset based on latitude, longitude, and presence of NaNs in the images has not significantly altered the distribution of flare classes.

2.6 Dataset Partitions

To facilitate comparison between flare prediction methods, we have partitioned the dataset into training, validation, and testing sets. To this end, we randomly selected 10% of the ARs to set aside for validation purposes (e.g., tuning of algorithm parameters), an additional 10% of the ARs for testing purposes, and the remaining 80% for training purposes. We note that the initial random assignment of ARs resulted in a validation set with different classification performance, specifically a higher true positive rate (TPR), on several classification tasks. Further investigation found that the validation set had a higher proportion of ARs with very high TPR. Randomly re-assigning seven ARs with TPR$>$0.90 from validation to test and a random seven ARs with TPR$<$0.90 from test to validation resulted in more similar performance between test and validation. There are 157 ARs and 94,757 images in the test data, 157 ARs and 95,933 images in the validation data, and 1,256 ARs and 759,357 images in the training data. Lists of the ARs included in each of the three sets are provided in files `List_of_AR_in_Train_Data_by_AR.csv`, `List_of_AR_in_Validation_Data_by_ARcsv`, and `List_of_AR_in_Test_Data_by_AR.csv` as part of the dataset repositories [15, 16].

2.7 Reduced Resolution Dataset

We have created a reduced resolution dataset at a spatial resolution of 224×224 pixels and bit depth of 8 bits. This reduced resolution dataset contains images in `.png` format which are more readily ingested by standard image processing libraries and at bit-depth and spatial resolution compatible with typical convolutional neural network (CNN) architectures. The spatial resolution is reduced using the `transform.resize` command in scikit-image with parameters `order=1, mode='reflect', clip=True, preserve_range=True, anti_aliasing=True`. The bit depth of the resized images are reduced by 1) offsetting the intensities by 2550, 2) clipping to the range $[0, 5100]$, 3) scaling to the range $[0, 255]$, and 4) rounding to the nearest integer:

$$I_8 = \left\lceil \text{MinMax}(0, 5100, I + 2550) \frac{255}{5100} \right\rfloor, \tag{1}$$

where I_8 is the image in unsigned 8-bit integer bit-depth resolution, I is the input image, $\text{MinMax}(mn, mx, x)$ denotes a clipping of x to the range $[mn, mx]$, and $[\cdot]$ denotes a rounding operation. It should be noted that this reduction in bit depth results in an error due to both the clipping operation and the scaling operation. The clipping operation affects only 2e-4% of pixels in the entire dataset which originally corresponded to the largest flux values (positive and negative). The scaling operation will result in a range of 20 G being mapped to the same intensity level with an error in the range $[-10, 10]$ G which is on the order of the noise level of the HMI instrument [21].

3 Data Records

The preconfigured dataset [15], reduced resolution dataset [16], and extra images dataset [17] each contain a directory structure `Lat60_Lon60_Nans0`, `Lat60_Lon60_Nans0_png_224`, and `active_regions_extra`, respectively. This directory structure contains the ARs in four digit directory names, e.g., `1325`. Each directory contains multiple magnetogram images in `.fits` format [15, 17] or `.png` format [16]. The base filenames are defined with the format `hmi.M_720s.YYYYMMDD_HHMMSS_TAI.1.magnetogram` as downloaded from JSOC.

The preconfigured dataset [15] and the reduced resolution dataset [16] additionally contain the following files of use for classification and regression tasks. In the following, the first filename corresponds to the preconfigured dataset [15] and the second filename corresponds to the reduced resolution dataset [16]; if only one filename is given, the filenames (and files) are identical between the two datasets.

- `C1.0_24hr_Labels.txt, C1.0_24hr_224_png_Labels.txt`: a file containing the labels for each of the images in the dataset. The labels are formatted to provide both the regression and classification labels in a form that can be parsed for other applications. Each line in the file is of the form `filename,label` where `filename` is the base filename in the image set and `label` is the label. The label is formatted as a string `KX.X` for flaring regions, where K is the GOES flare class (C, M, or X) and X.X is the size, e.g., `4.7`. Non-flaring regions are assigned a label of `'0'`.

- `List_of_AR_in_Train_Data_by_AR.csv, List_of_AR_in_Validation_Data_by_ARcsv`, and `List_of_AR_in_Test_Data_by_AR.csv`: files containing lists of NOAA ARs assigned to the training, validation, and test sets, respectively. Each line in the files is of the format `NNNN`, the four digit NOAA AR number. Note–these lists are identical between the reduced resolution dataset and the full resolution dataset.
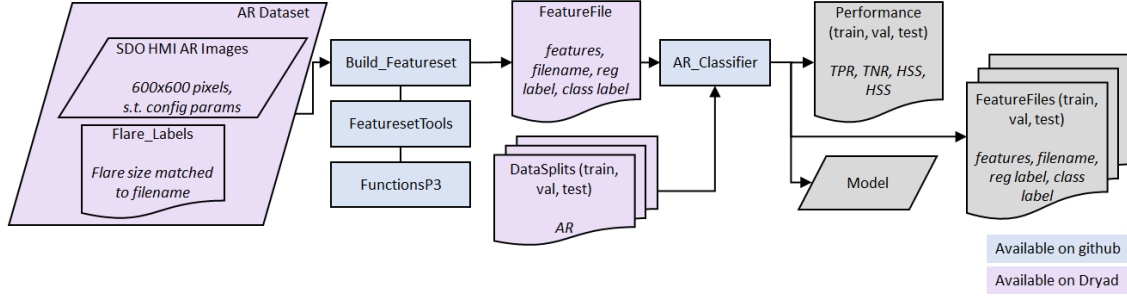
8

Figure 5: Flowchart of SVM classification of flare activity.

- `Lat60_Lon60_Nans0_C1.0_24hr_features.csv`, `Lat60_Lon60_Nans0_C1.0_24hr_png_224_features.csv`: a file with the 29 magnetic complexity features extracted from each of the images in the preconfigured datasets. Each line of the file contains 32 comma separated values. The first 29 values are the 29 magnetic complexity features as described below. The last three values are the classification label (1 or 0), regression label (flare size as as string `KX.X` or 0), and the base filename. The regression label is formatted as a string `KX.X` for flaring regions, where `K` is the GOES flare class (`C`, `M`, or `X`) and `X.X` is the size, e.g., `4.7`.
- (`Train_Data_by_AR.csv`, `Train_Data_by_AR_png_224.csv`), (`Validation_Data_by_AR.csv`, `Validation_Data_by_AR_png_224.csv`), (`Test_Data_by_AR.csv`, `Test_Data_by_AR_png_224.csv`): files with labels for each of the images in the preconfigured dataset formatted to provide classification labels in the format expected by a dataframe loader in tensorflow for the training, validation, and test sets, respectively. Each line is of the form `NNNN/filename,label` where `NNNN` is the AR directory, `filename` is the base filename, and `label` is the classification label (1 for flaring and 0 for nonflaring).

The extra images dataset [17] contains a file `EventList.txt` which contains the list of events (flares) occurring within the timespan of the dataset. Each line is of the format `YYYY MM DD,HHMM,NNNN,KX.X` where `YYYY MM DD` is the date, `HHMM` is the time, `NNNN` is the four-digit NOAA AR number, and `KX.X` is the GOES class (e.g., `C1.0` or `X10.1`).

4 Technical Validation

In this section we describe two experiments that demonstrate the utility of the preconfigured AR dataset. In the first, we implement a flare prediction method using magnetic complexity features and a support vector machine (SVM) classifier. In the second, we provide preliminary results of a transfer learning approach for use of convolutional neural networks (CNNs) for flare prediction.

4.1 Magnetic Complexity Features for Machine Learning

We extract 29 of the 38 magnetic complexity features of [5] from each of the HMI magnetograms in the preconfigured AR dataset and use a support vector machine (SVM) to predict whether the AR will flare within the next 24 hours. An overview of the SVM classification is shown in Figure 5. The methods presented in [5] were applied to MDI magnetograms which have lower spatial resolution ($\sim 2'' \times 2''$ pixels), and lower cadence (96 minutes) than the HMI dataset presented here ($\sim 0.5'' \times 0.5''$ pixels and 12 minute cadence). Due the lower cadence of the MDI magnetograms, the dataset was also much smaller, with approximately 260,000 total images. The 9 flux evolution features from [5] are omitted in this work: these features require a comparison between two images and therefore cannot be directly linked to a single image, the cadence of the HMI magnetograms is 12 minutes (as opposed to 96 minutes) leading to minimal evolution of an AR between images in this dataset, and the flux evolution features proved to be poor features for classifying ARs.

We provide the 29 magnetic features as part of the preconfigured AR dataset [15] and for the reduced resolution dataset [16] and the code to extract the magnetic features on GitHub at [10]. These magnetic complexity features include 7 gradient features, 13 neutral line features, 5 wavelet features, and 4 flux features in the format of a `.csv` file. Each row in the `.csv` file represents an image in the dataset. The first 29 columns are the 29 magnetic features. The 30th column is the binary flare class (`'1'` or `'0'`) and the 31st column is the flare size in terms of the GOES flare class (with a value of `'0'` representing no flare or a flare smaller than `'C1.0'`). The last column is the filename of the image corresponding to the magnetic features and flare class.

9

Method	Dataset	TPR	TNR	HSS	TSS
SVM	Full Resolution	0.7484	0.7791	0.4485	0.5275
	Reduced Resolution	0.7884	0.7464	0.4350	0.5348

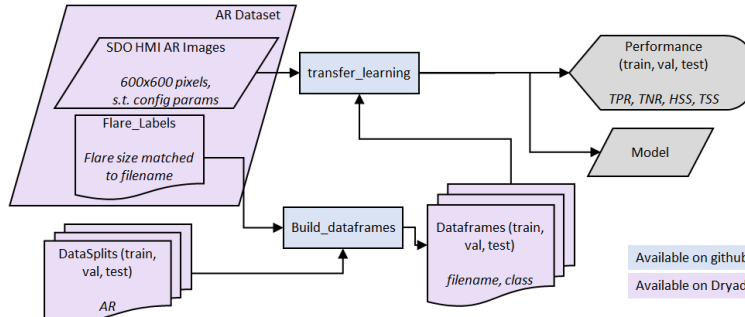Table 1: SVM performance on the test dataset for the full resolution and reduced resolution datasets.

Figure 6: Flowchart of VGG classification of flare activity.

An SVM classifier is trained on the training set using the `SVC` function from scikit-learn; this code is also available on GitHub at [10]. All parameters were left as the default (`C=1.0`, `shrinking=True`, `probability=False`, `tol=0.001`, `decision_function_shape='ovr'`, `break_ties=False`, `random_state=None`) with the exception of the `kernel` parameter which was set to `'linear'` and the `class_weight` parameter which was set to `'balanced'` to account for the imbalanced nature of this dataset. This experiment is intended as a validation of the use of the datasets for classical machine learning methods. As such, we have not optimized the kernel or parameters of the classifier. Performance metrics are evaluated on the test set and are summarized in Table 1. The performance metrics considered are the true positive rate (TPR), true negative rate (TNR), Heidke skill score (HSS), and true skill statistic (TSS) as defined in [5]. As a comparison the work in [5] achieved a TPR of 0.81, TNR 0.70, HSS 0.39, and TSS 0.51. Given that work was applied to a different dataset from a different instrument, we find the results here comparable to that work and a validation of the utility of this dataset for flare prediction. We also note that the comparable performance between the full and reduced resolution data indicates that the reduced resolution dataset has retained the vast majority of the information needed for this classification problem. We note, however, that other machine learning tasks may benefit from the increased spatial or bit depth resolution of the full resolution dataset.

4.2 Deep Learning

We perform transfer learning on the VGG16 CNN [22], pretrained on ImageNet using the `tensorflow.keras` [23] VGG model. An overview of the VGG classification is shown in Figure 6 and code is available on GitHub at [10]. We replace the final fully connected layer (originally 4096×1000) with a 4096×2 layer with softmax activation. In training, we freeze all layers except that final fully connected layer. For the full resolution data in `.fits` format, a custom data generator was written since the `.fits` format is not one that `tensorflow` can handle natively. Within that data generator, the images are resized to the expected spatial dimensions (224×224 pixels) using the `skimage.transforms.resize` command with options `order=1`, `mode='reflect'`, `clip=True`, `preserve_range=True`, and `anti_aliasing=True` and to the expected intensity range by linearly scaling the full range of the data $[-5978.7, 5978.7]$ to $[0, 255]$. The images are then preprocessed with the built-in `preprocess_input` function as part of the `tensorflow.keras` VGG model. For the reduced resolution dataset, the `flow_from_dataframe` method is used along with the VGG `preprocess_input` preprocessing. Both data generators use a batch size of 64. For training, we used the adam optimizer with options `learning_rate=0.001`, `beta_1=0.9`, `beta_2=0.999`, `epsilon=1e-07`, and `amsgrad=False` and the categorical cross-entropy loss. The networks are trained for 5 epochs with the `class_weight` parameter set to 1 for the majority (non-flare) class and N_n/N_f for the minority (flare) class, where N_n is the number of nonflaring examples and N_f is the number of flaring examples. We wrote custom `tensorflow.keras` metrics to track the TPR, TNR, HSS, and TSS throughout the training process. This experiment is intended as a validation of the use of the datasets for deep learning methods. As such, we have not optimized the architecture, which layers are frozen, or optimizer parameters. The best model was chosen as the epoch with the maximum validation TSS. Performance on the test data is summarized in Table 2. We see scores commensurate with the SVM performance, indicating the validity of this dataset in deep learning methods.

Method	Dataset	TPR	TNR	HSS	TSS
VGG	Full Resolution	0.6952	0.8142	0.4567	0.5094
	Reduced Resolution	0.7344	0.7980	0.4636	0.5325

Table 2: VGG performance on the test dataset for the full resolution and reduced resolution datasets.

Usage Notes

Further details on usage of the datasets can be found as part of the dataset repository documentation for the preconfigured dataset [15], reduced resolution dataset [16] and extra images dataset [17]. Further details on usage of the code for configuration of the datasets and classification can be found as part of the GitHub repository [10].

Code availability

All code used to generate and manipulate the dataset, as well as code used in the Technical Validation is available at the GitHub repository [10]. Further details and documentation regarding code usage are included therein.

References

[1] W. D. Pesnell, B. J. Thompson, and P. Chamberlin, "The solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1-2, pp. 3–15, 2012.

[2] J. R. Lemen, D. J. Akin, P. F. Boerner, C. Chou, J. F. Drake, D. W. Duncan, C. G. Edwards, F. M. Friedlaender, G. F. Heyman, N. E. Hurlburt, *et al.*, "The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1-2, pp. 17–40, 2012.

[3] T. Woods, F. Eparvier, R. Hock, A. Jones, D. Woodraska, D. Judge, L. Didkovsky, J. Lean, J. Mariska, H. Warren, *et al.*, "Extreme ultraviolet variability experiment (EVE) on the solar dynamics observatory (SDO): Overview of science objectives, instrument design, data products, and model developments," *Solar Physics*, vol. 275, no. 1-2, pp. 115–143, 2012.

[4] P. H. Scherrer, J. Schou, R. Bush, A. Kosovichev, R. Bogart, J. Hoeksema, Y. Liu, T. Duvall, J. Zhao, C. Schrijver, *et al.*, "The helioseismic and magnetic imager (HMI) investigation for the solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1-2, pp. 207–227, 2012.

[5] A. Al-Ghraibah, L. Boucheron, and R. T. J. McAteer, "An automated classification approach to ranking photospheric proxies of magnetic energy build-up," *Astronomy & Astrophysics*, vol. 579, p. A64, 2015.

[6] L. E. Boucheron, A. Al-Ghraibah, and R. T. J. McAteer, "Prediction of solar flare size and time-to-flare using support vector machine regression," *The Astrophysical Journal*, vol. 812, no. 1, p. 51, 2015.

[7] R. C. Canfield and P. Murdin, "Solar active regions," *Encyclopedia of Astronomy and Astrophysics*, vol. 3, pp. 2457–2462, 2000.

[8] "HMI observables data coverage plots." `http://jsoc.stanford.edu/data/cov.html`.

[9] "Space Weather Prediction Center SOLAR REGION SUMMARY (SRS)." `ftp://ftp.swpc.noaa.gov/pub/forecasts/SRS/README`.

[10] "AR-flares." *GitHub* `https://github.com/DuckDuckPig/AR-flares`.

[11] The SunPy Community, W. T. Barnes, M. G. Bobra, S. D. Christe, N. Freij, L. A. Hayes, J. Ireland, S. Mumford, D. Perez-Suarez, D. F. Ryan, A. Y. Shih, P. Chanda, K. Glogowski, R. Hewett, V. K. Hughitt, A. Hill, K. Hiware, A. Inglis, M. S. F. Kirk, S. Konge, J. P. Mason, S. A. Maloney, S. A. Murray, A. Panda, J. Park, T. M. D. Pereira, K. Reardon, S. Savage, B. M. Sipőcz, D. Stansby, Y. Jain, G. Taylor, T. Yadav, Rajul, and T. K. Dang, "The sunpy project: Open source development and status of the version 1.0 core package," *The Astrophysical Journal*, vol. 890, pp. 68–, 2020.

[12] "SolarSoft." `http://www.lmsal.com/solarsoft/sswdoc/index_menu.html`.

[13] "selenium 3.141.0." `https://pypi.org/project/selenium/`.

[14] "mozilla/geckodriver." `https://github.com/mozilla/geckodriver`.

[15] L. E. Boucheron, T. Vincent, J. A. Grajeda, and E. Wuest, "Active region magnetograms for solar flare prediction: Full resolution dataset." *Dryad* `https://doi.org/10.5061/dryad.dv41ns23n`, 2023.

[16] L. E. Boucheron, T. Vincent, J. A. Grajeda, and E. Wuest, "Active region magnetograms for solar flare prediction: Reduced resolution dataset." *Dryad* `https://doi.org/10.5061/dryad.jq2bvq898`, 2023.

[17] L. E. Boucheron, T. Vincent, J. A. Grajeda, and E. Wuest, "Active region magnetograms for solar flare prediction: Extra images dataset." *Dryad* `https://doi.org/10.5061/dryad.qjq2bvqmj`, 2023.

[18] "EDITED SOLAR EVENTS LISTS." `ftp://ftp.swpc.noaa.gov/pub/indices/events/README`.

[19] K. C. Fox, "Classifying solar eruptions." `https://www.nasa.gov/mission_pages/sunearth/news/classify-flares.html`, 2012.

[20] K. Leka, G. Barnes, and E. Wagner, "Evaluating (and improving) estimates of the solar radial magnetic field component from line-of-sight magnetograms," *Solar Physics*, vol. 292, no. 2, p. 36, 2017.

[21] S. Couvidat, J. Schou, J. T. Hoeksema, R. S. Bogart, R. I. Bush, T. L. Duvall, Y. Liu, A. A. Norton, and P. H. Scherrer, "Observables processing for the Helioseismic and Magnetic Imager instrument on the Solar Dynamics Observatory," *Solar Physics*, vol. 291, pp. 1887–1938, 2016.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." Preprint at `https://arxiv.org/abs/1409.1556`, 2014.

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from `https://www.tensorflow.org`.