Athina Bikaki and Christoph F. Eick

COSC 3337-17758: Data Science I
Fall 2025

**Helios: Analyzing and Understanding Solar Flares**



**Group Task 3** (260 pts)

Last Updated: Oct. 8 at 9a; due Date: Sa., Nov. 8 end of the day

Responsible TA: Athina Bikaki (abikaki@cougarnet.uh.edu)
Office hours: MO 2-3p/ TH 8:30-9:30a (MS Teams)

**A brief introduction to solar flares**

A **solar flare** is a sudden, intense burst of electromagnetic radiation originating from the Sun’s atmosphere [1]. Some characteristics of solar flares are:

* Accompanied by mass energy release
* Show a rapid increase in X-ray flux
* Typically, they last from a few minutes to several hours, and show a distinct temporal profile (rise and decay)
* Solar flares typically originate in the active regions of the Sun, particularly near sunspots, presenting opportunities for spatial analysis

Solar flares are categorized based on their X-ray brightness in the wavelength range of 1 to 8 angstroms, as measured by the Geostationary Operational Environmental Satellite (GOES), into the following classes:

1. A (weakest, often undetectable)
2. B (still weak)
3. C (small, common flares)
4. M (moderate, can cause radio blackouts)
5. X (extreme, rare)

Each letter represents a tenfold increase in energy. Solar flares can cause problems in communication, such as satellite damage, radio blackouts, and auroras.

🔭 In this project, we will work with **Solar Cycle 24**, the most recently completed solar cycle (December 2008 – December 2019), the 24th since 1755. Solar flare activity was minimal until early 2010. It reached its maximum in April 2014 [2]. We will use data from the Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI). RHESSI was a NASA solar flare observatory. It operated from February 2002 to August 2018. Unfortunately, RHESSI data are available until March 2018. Data were augmented with GOES observations during the same period. A snapshot of the data is shown below



In the following table, a brief explanation of the columns is provided.

|  |  |
| --- | --- |
| flare\_id | flare id, a column serving as a unique identifier  |
| date | date the solar event occurred |
| start\_time | start time of the event |
| peak\_time | peak time of the event |
| end\_time | end time of the event |
| duration | duration of the solar event in seconds  |
| peak\_counts | peak counts/second in energy range 6-12 keV, averaged over active collimators, including background |
| total\_counts | counts in energy range 6-12 keV integrated over the duration of the flare, summed over all sub-collimators, including background |
| energy\_band | the highest energy band in which the flare was observed |
| X\_pos | X position of the solar flare in arcseconds |
| Y\_pos | Y position of the solar flare in arcseconds |
| radial | distance from the Sun's center in arcseconds |
| AR | active region |
| flags | quality codes |
| SESC\_sunspot\_number | GOES SESC sunspot number (a column serving as a unique reference) |
| sunspot\_area | GOES sunspot area in 10 micro-hemispheres (μHem) |
| class\_str | GOES flare classification to C, M, and X. The value NF was introduced to denote no flare classification |

There is noise in this dataset in terms of class labels; not all solar flares have been matched, so several rows marked as 'no flare' (NF) could potentially be solar flares.

**Activity 1: Compile a solar flare dataset** (20 pts)

Process the dataset provided and answer the following questions:

1. Remove rows with zero position and energy range equal to 3-6 keV (1 pt)
2. What is the median duration in seconds of each solar flare class? (3 pt)
3. What is the average number of total counts of each solar flare class? (3 pt)
4. What is the average number of total counts yearly? Which year(s) were most active?
(3 pt)
5. During the most active year you found in (e), which month(s) were the most active?
(3 pt)
6. Which flags were mostly present in each solar flare class that occurred? (5 pts)
7. Add any new columns that could improve flare classification, and feel free to remove any that are redundant or not useful. (2 pts)

**Activity 2: Build a flare classifier** (40 pts)

Please use the dataset you have created in Activity 1 and build a machine learning model to learn to categorize solar flares (in classes C, M, X, or NF)

1. Create a new column named “class” with numeric labels from the column “class\_str”. Pre-process the data (e.g., normalization, fill missing values, numerical encoding of columns) (6 pts)
2. Split the dataset into train and test sets. Consider a way to split the data that will help you predict the flare’s class. Not all periods are equally active. The test set should not be used in model building and training. (3 pts)
3. Train a small (shallow) decision tree classifier with at most 30 nodes (leaves count as nodes). Try to find good hyperparameters in a cross-validation setting. Evaluate the classifier on the test data. (10 pts)
4. Report (i) the average training performance and the standard deviation in the cross-validation setting, and (ii) test performance in the test set using the evaluation metrics of your choice (e.g., accuracy, F1-score). Additionally, provide the confusion matrix. (6 pts)
5. Analyze the decision tree you created in step c. Summarize when the tree picks class M and when the classifier does not pick class M! What does this tell you about the characteristics of moderate flares? What about class X. Is there anything to learn from the tree about the characteristics of class X? (15 pts)

**Activity 3: RHESSI image cube exploration[[1]](#footnote-1)** (40 pts)

In this activity, you will use the two RHESSI solar flare images provided:

1. one representing an M-class solar flare on 2011-11-03
2. one representing a non-solar flare event on 2017-09-19

The selected RHESSI image cube files have 4 dimensions, with the first two dimensions representing the pixel values at X, Y positions, the third dimension corresponding to the energy values in each energy band, and the fourth dimension representing the time (X, Y, energy, time). The energy bands recorded, and the flare\_ids of the two images are:

1. 6-12, and 12-25 keV (ID: 11110332)
2. 3-6, 6-12, and 12-25 keV (ID: 17091901)

You can find these records in the dataset provided in Activity 1 by looking for the flare\_ids. Analyze these images to create the following comparative visualizations:

1. Visualize all energy bands for a single time frame – this is an energy band comparison (8 pts)
2. Visualize a selected energy band over time (8 pts)
3. Visualize all energy bands averaged over time (8 pts)
4. Create an animation over time and save it in a GIF or MP4 format (8 pts)
5. Pinpoint flare location over the Sun (8 pts)

For this activity, you will need to read astronomical data in FITS format. Both Python and R provide packages to read and process these files. Please refer to the useful resources at the end of this document for additional information.

*This activity does not depend on the dataset you created in Activity 1 and can proceed independently.*

**Activity 4: Temporal analysis** (40 pts)

In this activity, we will try to answer the following question: **How does flare intensity vary over time in a solar cycle?** Temporal analysis can uncover patterns tied to the 11-year solar cycle. For this activity, please use the dataset you have created in Activity 1.

1. Visualize flare intensity (use the column “total\_counts”) over time (8 pts)
2. Define a rolling time window (e.g., daily) and compute statistics (mean, std) of flare intensity over time (8 pts)
3. Visualize the mean over time (8 pts)
4. Visualize the std over time (8 pts)
5. How can we detect sudden changes in the data? Create a visualization that includes the flare intensity over time (a), the rolling mean (c), the rolling std (d), and annotate the points or windows you discovered as sudden changes (8 pts)

**Activity 5: Flare intensity assessment, mapping, and spatial analysis** (40 pts)

In this activity, we will try to answer the following question: **How do two different approaches compare in modeling the intensity and spatial distribution of solar flares?** Develop two methods for estimating flare intensity. The intensity estimation techniques measure the flare intensity at a specific location (X\_pos, Y\_pos) based on a set of flare events. Method 1 measures the intensity based on the “total\_counts” attribute. Method 2 measures the intensity based on the “duration” and “energy\_band”attributes.

1. Develop Method 1 (8 pts)
2. Develop Method 2 (8 pts)
3. Create intensity maps using the dataset you created in Activity 1 using Method 1 and Method 2 (14 pts)
4. Compare the two maps you generated: summarize the spatial variation and total intensity in each map (10 pts)

Remember that “energy\_band” **is summarized as bands, and there are 7 bands** intotal (3-6 keV was excluded). Creative visualization techniques will be granted more points.

**Activity 6: Story writing** (80 pts)

Create a PDF document with your findings, following the structure provided. There are no restrictions on the length of the document.

1. Introduction (8 pts)
* Define solar flares and their significance in space weather
* Brief overview of flare classification systems
* Motivation for analyzing flare data: prediction, impact, and solar cycle studies
* Summarize your contributions

B. Datasets (8 pts)

* Description of datasets used
* Pre-processing steps

C. Methodology (24 pts)

* Describe methods used for Activities 2,3,4, and 5

D. Results and analysis (24 pts)

* Explain your results/ findings
* Add your visualizations

E. Discussion (8 pts)

* Interpret your results
* Discuss any observations made

F. Conclusion (8 pts)

* A summary of key findings

**Submission requirements**

You may use any programming language or tool of your choice (e.g., Python, Excel, R) to complete the work. Please submit the PDF document along with a zip file that includes the source code, images, and animations you have created. Only one submission per group is required.

**Late submission policy**

Please check the late submission policy at <https://www2.cs.uh.edu/~ceick/UDM/COSC3337.html>

**Useful resources**

An example of how to load FITS (Flexible Image Transport System) image data in Python using the astropy package



A similar example in R (code translated using AI) using the FITSio package



A similar, more advanced example in R (code translated using AI) using the astro package



**References**

[1] “Solar flare,” *Wikipedia*. Sept. 26, 2025. Accessed: Sept. 27, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Solar\_flare&oldid=1313537737

[2] “Solar cycle 24,” *Wikipedia*. Sept. 03, 2025. Accessed: Sept. 27, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Solar\_cycle\_24&oldid=1309365374

1. Two examples of such images are provided in the specification. [↑](#footnote-ref-1)