# Topological Data Analysis Detects and Classifies Sunspots (Student Abstract)

**Aidan Lytle**[1], **Neil Pritchard**[1], **Alicia Aarnio**[1], **Thomas Weighill**[1]

[1]University of North Carolina at Greensboro
1600 Spring Garden St
Greensboro, North Carolina, 27406 USA
aflytle@uncg.edu

## Abstract

In our technology-dependent modern world, it is imperative to monitor the Sun for space weather threats to critical infrastructure. Topological data analysis (TDA) is a new set of mathematical techniques used in data analysis and machine learning. We demonstrate that TDA can robustly detect and classify solar surface and coronal activity. This technique is a promising step toward future application in predictive space weather modeling.

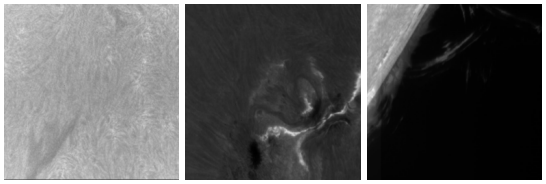
## Introduction

In the modern world, nearly all of our society uses electronics which are sensitive to solar activity. Because of that, there is a need to develop new detection, classification, and prediction methods for solar activity. Topological data analysis (TDA) is a mathematical approach to data analysis, which uses theory from algebraic topology to produce information about datasets and their structure. TDA has been applied to a variety of data types including geospatial data (Feng, Hickok, and Porter 2022), time-series data (Ravishanker and Chen 2019) and texture data. In this paper, we propose a new method inspired by the texture classifier in (Chung, Hull, and Lawson 2020) for detection and classification of sunspots. Our method uses the $0$- and $1$-dimensional persistent homology and persistence curves.

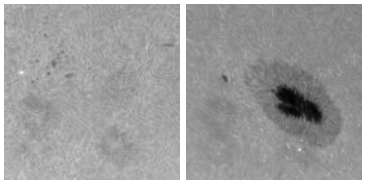
TDA is very good at finding data features that are traditionally thought of as "geometric". It is extremely good at detecting "holes" in data, in multiple dimensions. This is a powerful tool for solar activity, as, for example, sunspots are generally a form of "hole" in an image of the sun (see Figure 2). More specifically, we demonstrate that the 1st homology, which detects and classifies holes in 2-dimensional data, is an extremely powerful tool for sunspot detection and solar activity classification.

## Data

The dataset we worked with was the set used by Armstrong and Fletcher in (Armstrong and Fletcher 2019). This is a broad set of 13159 H$\alpha$ images of the sun classified into 5 categories. The categories are determined by well understood classes of solar surface activity, and are spots, flares, filaments, prominences, and quiet sun (see Figure 1. The images are broken into smaller sub-images to speed computation and improve accuracy; this was performed by Armstrong. This is imagery drawn from the *Hinode/Solar Optical Telescope* (SOT) 2006 set. H$\alpha$ data is abundant and the amount of data available is on the order of petabytes.



(a) Filament (b) Flare (c) Prominence



(d) Quiet sun (e) Sunspot

Figure 1: Classes of solar activity.

## Methods

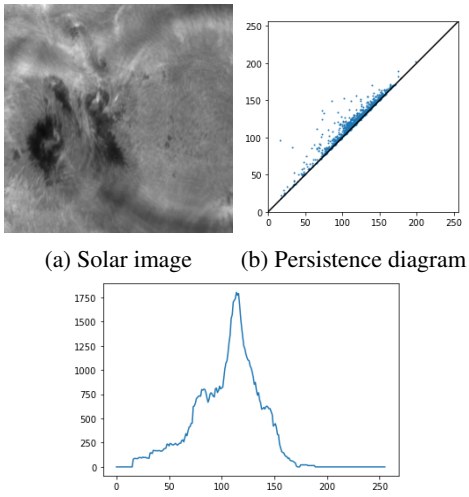
### Persistent Homology

The main tool in TDA is persistent homology. Persistent homology uses algebraic methods to track the appearance and disappearance of holes in a sequence of related spaces.

The method for detecting these computationally depends on the type of structure desired, but the most common method for image processing is sublevel set filtration (Chung, Hull, and Lawson 2020; Dunaeva et al. 2016).

Sublevel set filtrations are constructed from simplicial or cubical complexes equipped with a function on the simplices. For a given threshold $T$ , we define the sublevel set of an image to be all the pixels of brightness value less than $T$. These create a binary pixel map for each T from which the topological information is gained. Then for $S < T$ we have an inclusion map, which constructs a sequence of expanding pixel regions. We track connected components and holes

across this sequence of binary images, observing when they appear and when they disappear (by merging with another component, or in the case of a hole, by being closed up). This gives us birth and death values for each feature. These births and deaths are plotted against each other to form a persistence diagram. The resulting diagrams are then used to construct persistence curves. It should be clear that the points in these plots always satisfy $(d \geq b)$, and so they live above the line with slope 1. Figure 2 shows a solar image and its 0-dimensional persistence diagram.



(a) Solar image (b) Persistence diagram

(c) The persistence curve

Figure 2: The topological signature of the image is given by the diagram, and vectorized as the curve.

## Persistence Curves

After producing the persistence diagrams, we vectorize them using persistence curves. The persistence curve is a mapping

$$\psi : \{(b, d) \mid (b, d) \in \mathbb{R}^2, 0 \leq b \leq d\} \rightarrow L^2(\mathbb{R})$$

We take an arbitrary parameter $t \in [0, max(d)]$, and use the "lifetime" of each point in the box outlined by $b < t, d > t$. We denote this box $B(t)$. We then define the curve of diagram $D$ with parameter $t$ to be

$$\psi_D(t) = \sum_{(b,d) \in B(t)} d - b$$

This returns a curve, which is treated as a vector in $L^2$. Classification can now be performed by standard machine learning algorithms. We in this study utilize a support vector machine (SVM).

## Results

We test our method on 1318 test images with a 80-20 train-test split, producing a 97% accuracy ( Figure 3). This is compared to the deep convolutional network utilized in (Armstrong and Fletcher 2019), which had accuracy $> 99.9\%$.

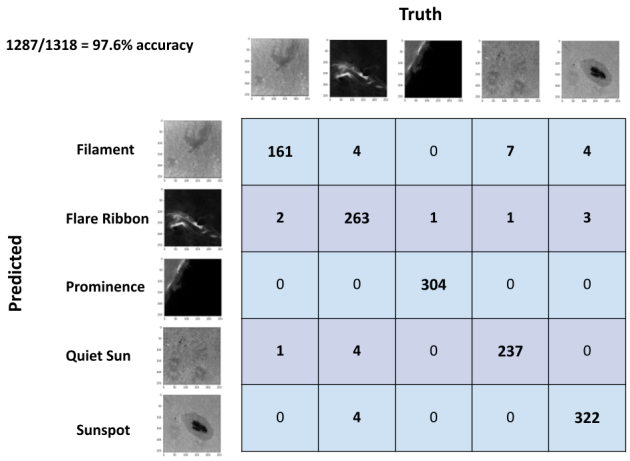



Figure 3: Confusion matrix for the classification using persistent homology and persistence curves. Our method obtained an accuracy of 97.6%.

## Discussion

The results of this test were strong, in that the algorithm was highly accurate. Notably, our topological features are theoretically invariant under global shifts in brightness as well as rotations. We expect this to provide additional robustness across datasets. The method is a conceptual improvement on most deep learning frameworks, in the sense that it uses more explainable features – namely, holes in the image data. This trade off comes at the cost of reduced accuracy, though accuracy can be improved by adding additional topological features. Our study found an accuracy of 98.5% with the combined data from the 1-dimensional and 0 dimensional homology (the curves are concatenated), and 98.8% with the usage of combined 1-dimensional and 0 dimensional homology and the image inverses. Future work could include a time-series topological analysis to aid in real-time prediction.

## References


Armstrong, J. A.; and Fletcher, L. 2019. Fast solar image classification using deep learning and its importance for automation in solar physics. *Solar Physics*, 294(6): 1–23.

Chung, Y.-M.; Hull, M.; and Lawson, A. 2020. Smooth summaries of persistence diagrams and texture classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 840–841.

Dunaeva, O.; Edelsbrunner, H.; Lukyanov, A.; Machin, M.; Malkova, D.; Kuvaev, R.; and Kashin, S. 2016. The classification of endoscopy images with persistent homology. *Pattern Recognition Letters*, 83: 13–22.

Feng, M.; Hickok, A.; and Porter, M. A. 2022. Topological data analysis of spatial systems. In *Higher-Order Systems*, 389–399. Springer.

Ravishanker, N.; and Chen, R. 2019. Topological data analysis (TDA) for time series. *arXiv preprint arXiv:1909.10604*.