Identification of Coronal Holes on AIA/SDO images using unsupervised Machine Learning

Fadil Inceoglu [ID],[1] Yuri Y. Shprits,[1,2,3] Stephan G. Heinemann,[4] and Stefano Bianco[1]

[1]*GFZ German Research Centre for Geosciences, Potsdam, Germany*

[2]*Institute for Physics and Astronomy, University of Potsdam, Potsdam, Germany*

[3]*Department of Earth, Planetary, and Space Science, University of California, Los Angeles, Los Angeles, CA, USA*

[4]*Max-Planck-Institute for Solar System Research, Goettingen, Germany*

ABSTRACT

Through its magnetic activity, the Sun governs the conditions in Earth's vicinity, creating space weather events, which have drastic effects on our space- and ground-based technology. One of the most important solar magnetic features creating the space weather is the solar wind, that originates from the coronal holes (CHs). The identification of the CHs on the Sun as one of the source regions of the solar wind is therefore crucial to achieve predictive capabilities. In this study, we used an unsupervised machine learning method, k-means, to pixel-wise cluster the passband images of the Sun taken by the Atmospheric Imaging Assembly on *the Solar Dynamics Observatory* (AIA/SDO) in 171 Å, 193 Å, and 211 Å in different combinations. Our results show that the pixel-wise k-means clustering together with systematic pre- and post-

Corresponding author: Fadil Inceoglu
fadil@gfz-potsdam.de

processing steps provides compatible results with those from complex methods, such as CNNs. More importantly, our study shows that there is a need for a CH database that a consensus about the CH boundaries are reached by observers independently. This database then can be used as the "ground truth", when using a supervised method or just to evaluate the goodness of the models.

Keywords: coronal holes, k-means, machine learning

1. INTRODUCTION

The Sun is a magnetically active star that shows various magnetic activity structures extending from its surface to its higher atmospheric layers, such as bipolar active regions (ARs) on the photosphere, filaments in the chromosphere, and coronal holes (CHs) in its corona. Through its magnetic activity, the Sun governs the conditions in the vicinity of Earth and throughout the heliosphere, which creates space weather and space climate. Space weather is defined as the effects of the solar wind, and solar eruptive phenomena, such as flares and coronal mass ejections (CMEs), on Earth's magnetosphere, ionosphere, and thermosphere (Schwenn 2006). The space weather conditions have drastic effects our space- and ground-based technology (Eastwood et al. 2017).

One of the most important solar magnetic features creating the space weather and in turn affecting the Earth, is the solar wind. The observations revealed that there are three different types of solar wind; (i) steady fast solar winds originate in the CHs, (ii) unsteady slow winds from opening magnetic loops and active regions, and (iii) transient winds from CMEs (Marsch 2006). The identification of the CHs on the Sun as one of the source regions of the solar wind (Wilcox 1968) that creates space weather and in turn influences our space- and ground-based technology is therefore crucial to achieve predictive capabilities.

As the source regions of the steady fast solar winds, CHs are identified as regions of low density collisionless plasma that is generally located above inactive parts of the Sun, where open magnetic field lines extend throughout the heliosphere (Schwenn 2006; Cranmer 2009). The magnetic field inside a CH is known to be more unipolar and the CHs show sharp and/or diffuse transition on the

boundaries between them and their surroundings (Cranmer 2009). The temporal evolution of the CH as well as the area they cover on the Sun depends on the solar activity cycle, also known as the Schwabe cycle (Schwabe 1844). During the minimum phase of a solar cycle, the CHs are observed to be larger and located mainly on the solar polar caps. On the inclining phase of a cycle, the CHs are observed to be present at any latitude and to be short-lived. During solar maximum, the CHs are smaller and only exist around mid-latitudes, while on the declining phase of the solar cycle there are more long-lived CHs at lower latitudes and they form closer to the solar equator as the cycle progresses (Hewins et al. 2020). Additionally, during the inclining and declining phases of a solar cycle, the CHs can evolve into structures extending from a solar pole to solar equator.

As CHs have lower densities and temperatures, and hence the lowest emission in UV and X-ray in comparison to their surrounding environment consisting of active regions and quiet Sun, they appear as dark regions solar images in wavelengths around 194 Å whether they are on-disk or off-limb CHs (Cranmer 2009).

Detection of CHs are done by eye-based on He I 10830 Å near-infrared absorption line triplet (Harvey & Recely 2002), histogram-based intensity thresholding on 193 Å and 195 Å passband images of the Sun from the Atmospheric Imaging Assembly (AIA; Lemen et al. 2012) on *the Solar Dynamics Observatory* (SDO; Pesnell et al. 2012) and the Extreme Ultraviolet Imaging Telescope (EIT; Delaboudinière et al. 1995) on *the Solar and Heliospheric Observatory* (SOHO), respectively (CHARM; Krista & Gallagher 2009). Additionally, an automated method for detection and segmentation of CHs based on multi-thermal intensity segmentation using 171 Å, 193 Å, and 211 Å passband images of the Sun from the AIA/SDO (CHIMERA; Garton et al. 2018), and semi-automated method based on intensity threshold that is modulated by the intensity gradient of a CH have been developed (CATCH; Heinemann et al. 2019).

There are also methods based on supervised and unsupervised machine learning (ML) methods. Verbeeck et al. (2014) developed a set of segmentation procedures based on spatial possibilistic clustering algorithm (SPoCA) to detect CHs in an unsupervised ML fashion. Identified ARs and CHs by this algorithm are uploaded to the event catalogs in the Heliophysics Event Knowledge (HEK)

database (Hurlburt et al. 2012). Illarionov & Tlatov (2018) used convolutional neural networks (CNNs; Schmidhuber 2014; Lecun et al. 2015) based on the U-net architecture (Ronneberger et al. 2015) to identify CHs on solar images at 193 Å passband images of the Sun from AIA/SDO. They trained their network using binary maps from Kislovodsk Mountain Astronomical Station. Recently, Jarolim et al. (2021) utilized a progressively growing architecture based CNNs using data from all 7 channels of AIA/SDO (94 Å, 131 Å, 171 Å, 193 Å, 211 Å, 304 Å and 335Å) as well as line-of-sight magnetograms from Helioseismic and Magnetic Imager (HMI; Scherrer et al. 2012) on the SDO. For their network, the authors used binary maps from manually reviewed SPoCA-CH (Delouille et al. 2018).

In this study, we utilize pixel-wise k-means algorithm, which is an unsupervised ML method, to detect CHs based on 171 Å, 193 Å, and 211 Å passband images from the AIA/SDO. To achieve this objective, we used data from each channel in different combinations, and compared results from each combination to each other as well as to those from CATCH and the HEK data to calculate their performances. We first describe the data used in this study in Section 2 and explain the analyses and present our results in Section 3. We discuss the results and conclude in Section 4.

2. DATA

To detect the CHs on the solar corona, we use passband data with 2 second exposure from AIA/SDO in wavelengths 171 Å, 193 Å, and 211 Å in different combinations (Figure 1). The AIA telescope on the SDO takes passband measurements of the Sun in every 12 seconds in full disk with a spatial resolution of 4096×4096 pixels, and each pixel corresponds to 0.6 arsec on the solar disk leading to a spatial resolution of 1.5 arcsec (Lemen et al. 2012). These 3 EUV bandpasses are centred on specific spectral emission lines of Fe IX for 171 Å, Fe XII, XXIV for 193 Å, and Fe XIV for 211 Å, which covers the temperature range from 6×10^5 to 2×10^6 K, corresponding to the upper transition region, quiet corona (171 Å), corona and hot flare plasma (193 Å), and active-region corona (211 Å) (Lemen et al. 2012).
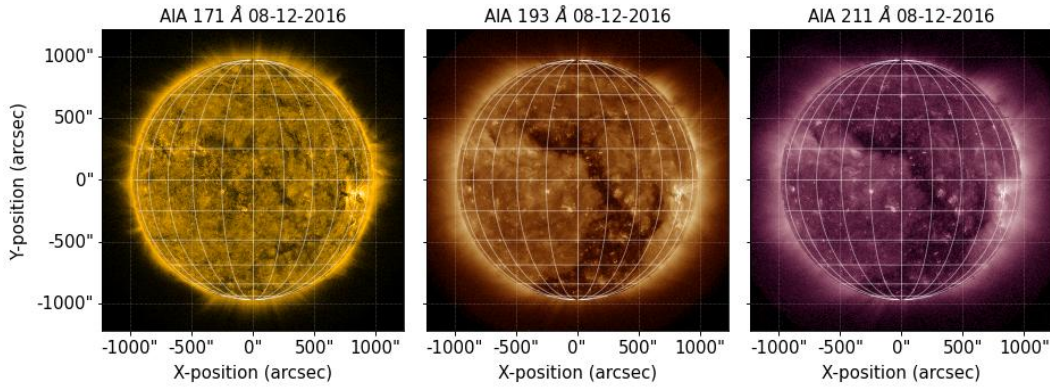
3. ANALYSES AND RESULTS

Figure 1. Passband images of the Sun in 171 Å (the left panel), 193 Å (the middle panel), and 211 Å (the right panel) taken by the AIA/SDO on 8 December 2016 at 00:00 UT.

3.1. *Preprocessing data*

To detect the CHs, we use solar images taken by AIA/SDO in passband images in wavelengths 171 Å, 193 Å, and 211 Å in different configurations. We also study the most efficient wavelength or configuration of wavelengths to identify the CHs. To achieve this, we compare our CH binary maps with those from the CATCH . We also compared the CH polygons provided by the HEK with the CATCH binary maps to have a base-line with which we compare our results. The CATCH binary maps are selected from the last two months of each year in a time-range from November 2010 to December 2016, extending through solar cycle 24. The CATCH data in this period is reliable with minimal uncertainties. The total of 237 CATCH CH binary maps consist only contributions from the longitudinal range of $[-400, 400]$ arcseconds in helioprojective coordinates as in this region the CHs can be identified more robustly (Jarolim et al. 2021). We also imported CH polygons from the HEK database for the same dates as the CATCH maps, and converted them into binary maps.

In total, we analyze 237 days of data. for each date, we import the level 1 data in 171 Å, 193 Å, and 211 Å wavelengths and preprocess them using *aiapy* (Barnes et al. 2020a,b) and *SunPy* (The SunPy Community et al. 2020; Mumford et al. 2021) python packages. This step consists of correcting the data for instrument degradation, for pointing and observer location. Following to these corrections, we registered and aligned the data and normalize it so it has a unit of count/pixel/second. Following these corrections, we correct the passband images for limb brightening using annulus limb
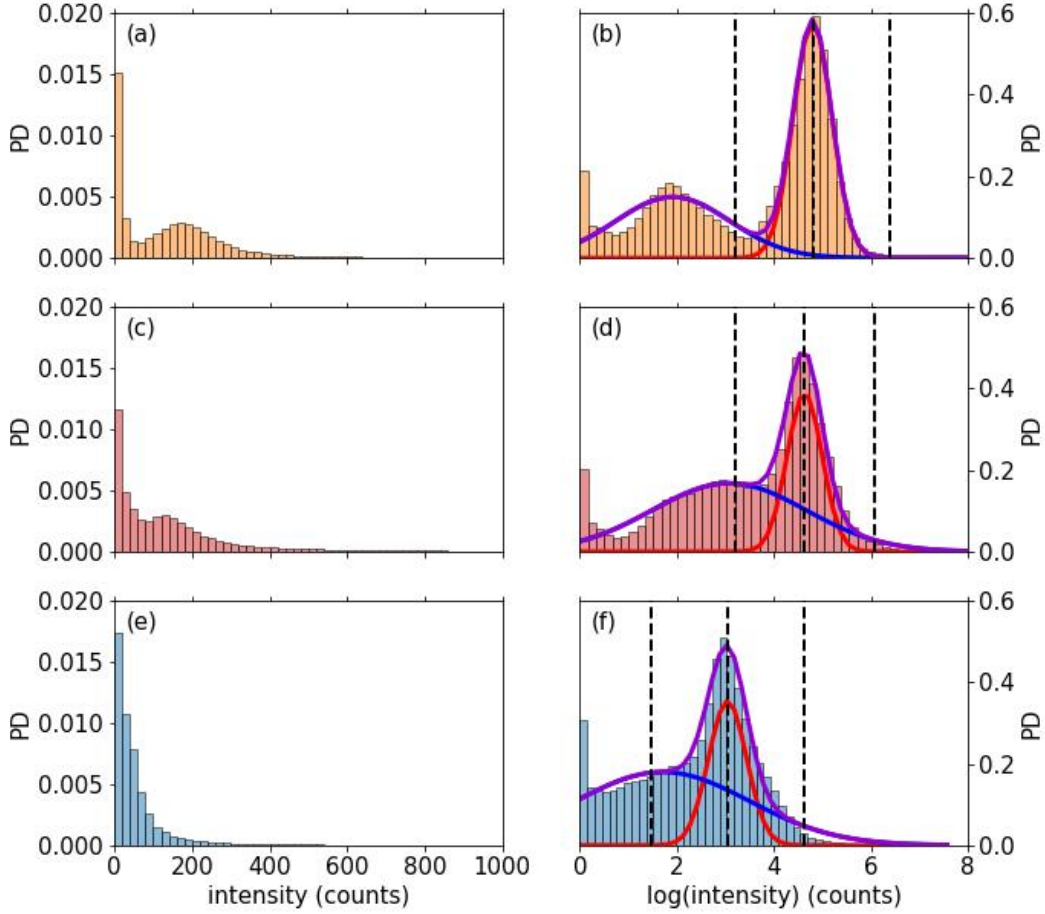
Figure 2. Probability densities of AIA/SDO 171 Å (top panel), 193 Å (middle panel), and 211 Å (bottom panel) intensities of the solar disk on 8 December 2016 at 00:00 UT. The left panels show the probability densities of the preprocessed data, while the right panels show probability densities of the post-processed data. The vertical dashed lines show mean (μ) and $\mu \pm 4\sigma$ values calculated to identify the threshold values.

brightening correction approach (Verbeeck et al. 2014). We then deconvolve the passband images using instrument point spread function for each wavelength, and rescaled them to 1024×1024 using spline method. As the final step, we log-norm transformed the data.

Following these steps, we created histograms of each data set to determine the lower and upper threshold values. Determining these values allows us to increase the contrast in the data. To avoid using any arbitrary values for these thresholds and to have a more systematic approach for determining these values, we fit a bimodal gaussian curve to each histogram (Figure 2), where it is possible. For some dates, however, it was not possible to fit a bimodal gaussian fit. For these dates, we used a

unimodal gaussian fit. Using the obtained parameters of the gaussian fits, we calculated the lower- and upper-threshold values based on the mean and standard deviation values of the higher peak (the right panels of Figure 2), because the lower peak represents the CH pixels (Heinemann et al. 2019). For each date in the dataset, we calculate a lower-threshold value for each wavelength based on $(\mu - 4\sigma)$, while the upper-threshold value is determined based on $(\mu + 4\sigma)$. Values below (above) the lower-threshold (upper-threshold) value are stacked to have only one value that is the threshold value.

We then investigate the temporal variations in the calculated mean (μ) and the lower threshold values $(\mu - 4\sigma)$ (Figure 3). The μ values of 193 Å and 211 Å passband images show variations in phase with the solar cycle, while the μ values of 171 Å does not show such a trend (Figure 3a). The μ values for each passband images also show day-to-day fluctuations. Similarly, the lower threshold values show day-to-day fluctuations as well. These fluctuations have wider range for the threshold values calculated for the 211 Å passband images especially during the maximum phase of the solar cycle, while the other two channels do not exhibit such wide fluctuations (Figure 3b). An important feature to note is the "negative" threshold values found for the 211 Å passband images. There are 27 days where the lower thresholds are negative values. However, as this does not have a physical meaning, the threshold values for these days were accepted as zero. The reason for the negative values come from the underlying shape of the gaussians.

3.2. *Pixel-wise clustering the images using the k-means algorithm*

After increasing the contrast in each image based on their individual mean and standard deviation values, we created 4 different data sets; (i) 193 Å image, (ii) 211 Å image, (iii) 193Å and 211Å composite image (2 channel composite, 2CC), and (iv) 171 Å, 193Å, and 211Å composite image (3 channel composite, 3CC). We then pixel-wise cluster each image using the k-means method. This method is used to automatically cluster a given data set into k groups of equal variance (MacQueen 1967). The most commonly used clustering criterion is the sum of squared Euclidian distances (SSD), also known as the within-cluster sum-of-squares, of each data point to centroid of the cluster, to which that data point is attained (Likas et al. 2003). The k-means algorithm first randomly
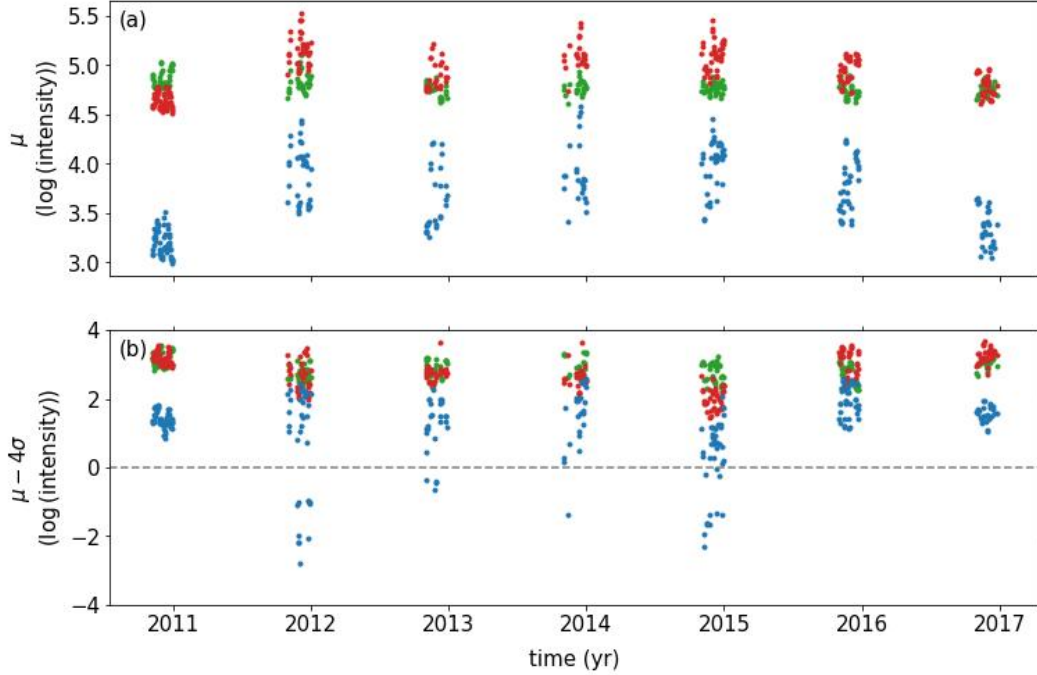
Figure 3. Calculated mean (μ) (a) and lower threshold values ($\mu - 4\sigma$) (b) for AIA/SDO 171 Å (green), 193 Å (red), and 211 Å (blue) passband images for the study period. Note that there are 27 points below zero, meaning that no lower threshold value could be calculated, therefore no thresholding applied to the 211 Å passband data on these dates.

selects k cluster centroids, and then iteratively refine these initial cluster centroids by assigning each Euclidian distance to its closest cluster centroid. Then the algorithm updates each cluster centroid value to be the mean of its elements by minimizing the SSD (Wagstaff et al. 2001; Likas et al. 2003).

The number of clusters, the k value, for this method is an input parameter. To choose the optimum number of clusters, we used the scree-plot method (Paparrizos & Gravano 2015). In this method, we use k = 1, 2, 3, ..., 10 and calculate the the sum of squared distances (SSD) for each k value. The results show that after the cluster number 3, any further decrease in SSD is very small compared to previous ones, which means that the optimum k value to use, is 3 (Figure 4). This indicates that there are darker regions, brighter regions, and regions that surround them, which can be attained to the CHs, active regions, and the quiet Sun.
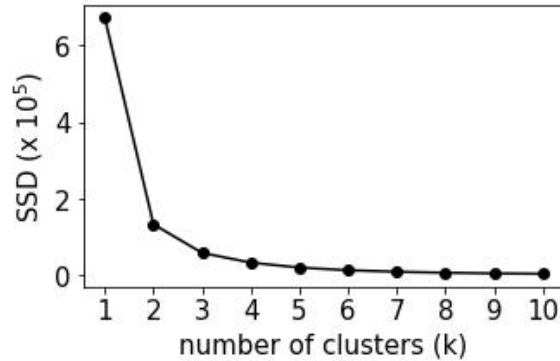
Figure 4. Sum of squared distances (SSD) calculated for each number of clusters, which ranges from 1 to 10 for passband data in 193 Å on 8 December 2017 at 00:00 UT.

The k-means method allows us to determine a threshold value for single channel inputs, a threshold line for 2 channel inputs, and a threshold surface for 3 channel inputs in a systematic way that enables us to deter from choosing these thresholds arbitrarily. Additionally, this method, when automated, is flexible enough for day-to-day variations in solar images, providing a dynamical response to them.

We calculate segmentation maps for each date using k-means method throughout solar cycle 24. Following that, we convert these maps to binary maps by merging the 2 clusters that identify brighter regions (active regions) and regions that surrounds darker and brighter regions (quiet sun). The reason we did not use k value as 2, is to avoid overestimation of the darker pixels on the passband images of the solar disk. We then remove small dotted-like regions using *morphology* module of scikit-image package (van der Walt et al. 2014). This method requires two inputs; the smallest allowable object size and connectivity, which we use 200 and 10 pixels, respectively. We also used morphological closing using a disk-shaped footprint with a radius of 2 pixels to remove smaller holes in identified CHs. The reason for using a smaller footprint is to try to avoid smoothing out larger bright points in identified CHs, which might be related to the Coronal Bright Points (Karachik et al. 2006; Hong et al. 2014; Wyper et al. 2018).

In addition to the 4 different binary maps types generated based on the 193 Å, 211 Å, 2CC, and 3CC, we generated another type of binary map. We generated them based on the overlap between binary maps of the 193 Å and 211 Å images, which we will refer to as the 2 Channel Overlap (2CO).

The 2CO binary maps are created if a pixel is simultaneously identified as a CH pixel in the two binary maps from the 193 Å and 211 Å images. Those pixel, which are not simultaneously identified as a CH are then accepted as non-CH pixels.

3.3. *Pixel-wise evaluation metrics*

To calculate the performances of our binary maps generated by the k-means method for each date, we used pixel-wise evaluation metrics. As there will be an imbalance between non-CH and CH pixels in the passband and composite images of the Sun, we use intersection over union (IoU), also known as the Jaccard index (Jaccard 1912), and true skill statistics (TSS) (Hanssen & Kuipers 1965) as pixel-wise evaluation metrics. To calculate these metrics, we used binary maps from CATCH. IoU and TSS are calculated based on each confusion matrix for each date using;

$$IoU = \frac{TP}{TP + FP + FN},$$ (1)

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN}$$ (2)

where TP, TN, FP, and FN denote pixel-wise calculated number of true positives, true negatives, false positives, and false negatives, respectively.

The distributions of the IoU values calculated between our and the CATCH binary maps together with those between the HEK and the CATCH binary maps show that the IoU for the HEK CH binary maps has a median value of 0.53±0.13, while our results from the AIA 193 and 2CC show median values of 0.62±0.14 and 0.64±0.14, respectively. This indicates a better overlap of the identified CHs from our method with those generated by CATCH. The other three binary maps from our study, the AIA 212, 3CC, and 2CO, result in IoU values of 0.51±0.20, 0.50±0.21, and 0.61±0.19, respectively (Figure 5a)..

The median TSS values of the AIA 193 and 2CC are 0.91±0.06 and 0.93±0.06, respectively (Figure 5b), while the median TSS value for the HEK is 0.73±0.13. These results indicate that our binary
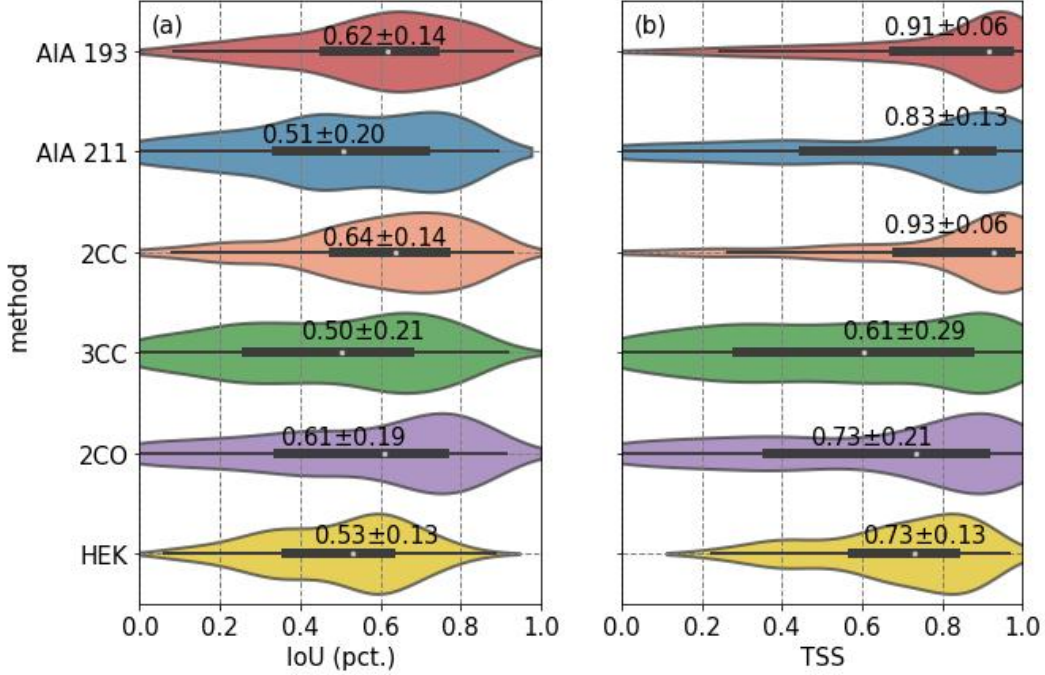
Figure 5. The distributions of the calculated IoU (a) and TSS (b) values between binary maps generated in this study and the CATCH, together with those between the HEK database and the CATCH. The white dots indicate the median value for each distribution. We also show the median values together with median absolute deviation for each evaluation metric in the figure. The red, blue, orange, green, purple, and yellow colors show AIA 193, AIA 211, 2CC, 3CC, 2CO, and HEK binary maps, respectively.

maps generated by AIA 193 and 2CC are more in line with those from CATCH. The AIA 212, 3CC, and 2CO, show median TSS values lower than AIA 193 and 2CC (Figure 5b).

3.4. *Coronal hole areas*

To further validate our results against the HEK and CATCH results, we calculate the total areas of the CHs on the solar disk in percentage of CH coverage on the solar disk. To achieve this, we first corrected each pixel in our binary maps for projection effects by applying;

$$A_i = \frac{A_{i,proj}}{\cos \alpha_i}, \tag{3}$$

where A_i and α_i denote the corrected pixel area and the heliographic angular distance of each pixel to the center of the solar disk as seen from the AIA/SDO, respectively.
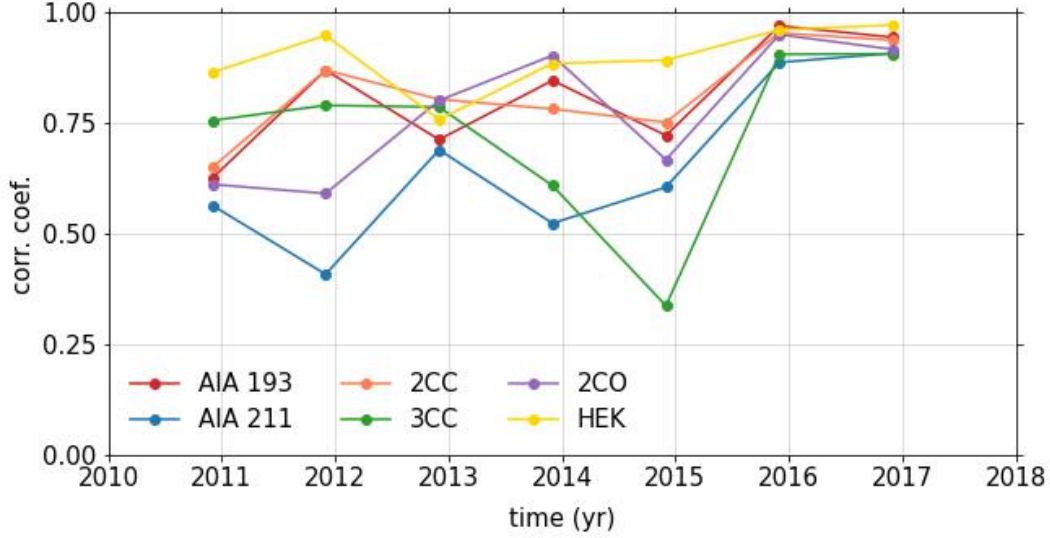
Figure 6. Temporal evolution of the correlation coefficients between total CH areas from our method, HEK against CATCH data through November 2010 and December 2016, extending through solar cycle 24. Note that the correlations are calculated using data during the last two months of each year (see text).

We calculated the Pearson correlation coefficients for each year between results from our study, HEK binary maps and CATCH (Figure 6). We need note that we use the last two months of each year to calculate the correlations. Similar to the results obtained for IoU and TSS, AIA 193 and 2CC generally provide higher correlations through the study period. Interestingly after 2014, the correlation coefficients calculated for every binary map become similar and evolve in parallel until 2016 (Figure 6).

We also calculated the overall correlations between the binary maps from our study and HEK, and binary maps from CATCH. The highest correlation of 0.88 for the CH areas is observed between the HEK and the CATCH data, while our 2CC gave a correlation coefficient of 0.82, followed closely by AIA 193 that gave a correlation coefficient of 0.81. The correlation coefficients for the 2CO, 3CC, and AIA 212 are 0.79, 0.75, and 0.73 respectively (Figure 7).

3.5. *Comparison of the CH binary maps*

We then select three dates that represent different phases of solar cycle 24 to compare the CH binary maps. These dates are (i) 05 November 2012 on the inclining phase before the cycle maximum, (ii)
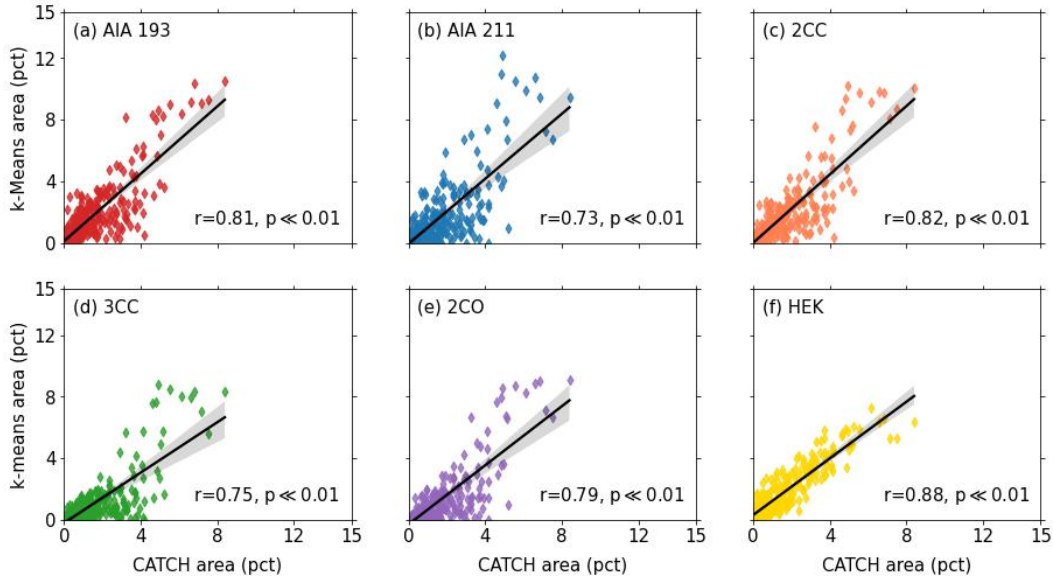
Figure 7. The total percentage areas from this study (a to e) and HEK data base (f) as a function of the areas from CATCH. The black solid lines show the linear fits, while the shaded areas show uncertainty. We also show the Pearson correlation coefficients and their statistical significances. The color coding is the same in Figure 5.

07 December 2014 right after the solar cycle maximum, and (iii) 07 December 2016 on the declining phase of solar cycle 24 (Figure 8).

On the inclining phase of solar cycle 24, on 05 November 2012, our method identifies smaller CHs. The results from the AIA 193, 3CC, and 2CO are observed to be more in line with those from the CATCH, where there is only one CH at $[0, 500]$ arcseconds in helioprojective coordinates. The results from the AIA 211 and the 2CC, on the other hand, more in line with those from the HEK database (the top row of Figure 8). On 07 December 2014, a few months after the cycle maximum, the binary maps from the AIA 193, the 3CC, and the 2CO show similar CH coverage on the solar disk to the CATCH within the longitudinal range of $[-400, 400]$ arcseconds. All of the CH binary maps from our method, except for the 3CC, are similar to the CHs from the HEK showing a small coronal hole near $[-750, 500]$ arcseconds (the middle row of Figure 8). On the declining phase of solar cycle 24, on 07 December 2016, the CH areas identified using the AIA 193, the 2CC, and the 3CC are in line with those from the HEK database and CATCH. On this date, the total CH area coverage also reaches
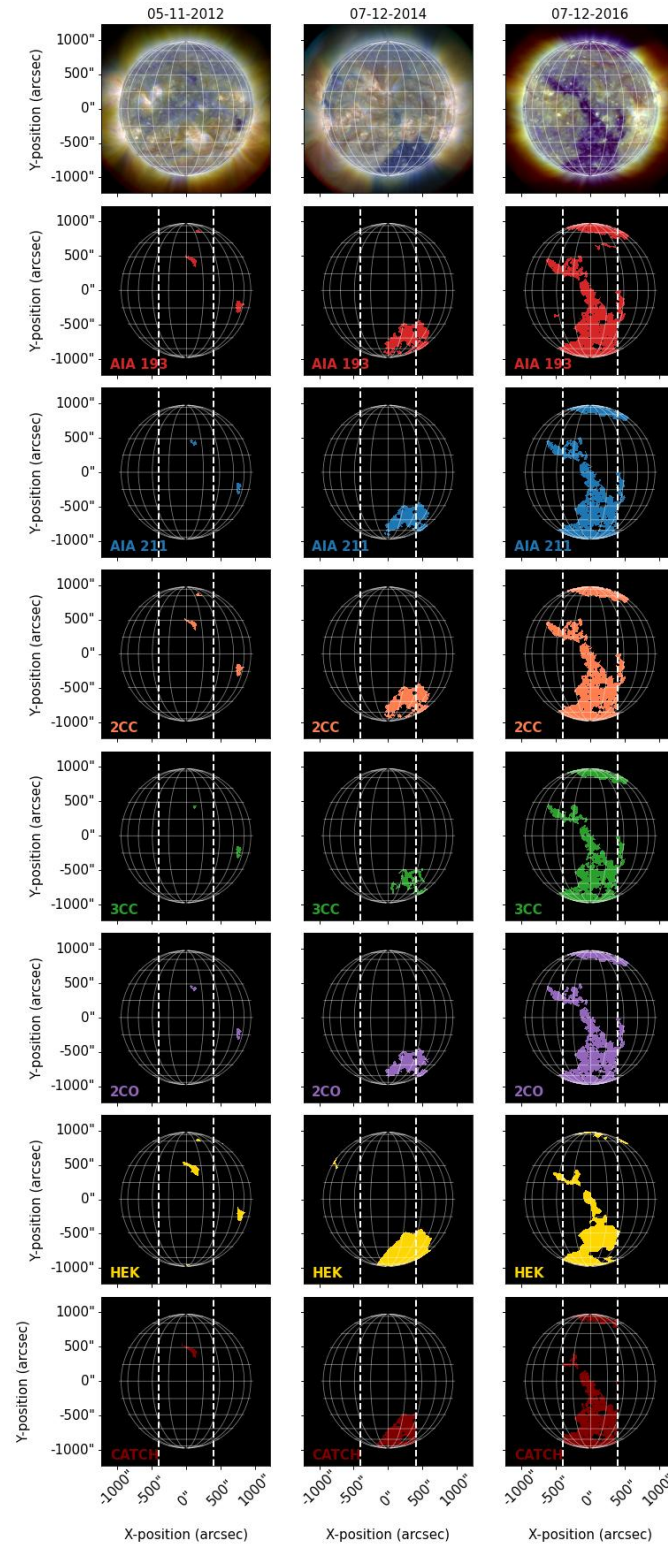
Figure 8. The CH binary maps for 05 November 2012 (top row), 07 December 2014 (middle row), and 07 December 2016 (bottom row) identified from the AIA 193, AIA 211, 2CC, 3CC, 2CO together with binary maps from the HEK and CATCH. The vertical white dashed lines indicate the longitudinal range of $[-400, 400]$ arcseconds in helioprojective coordinates. The color coding is the same in Figure 5.

its maximum, where it extends from the southern solar pole to the solar equator (the bottom row of Figure 8).

To evaluate the consistency of our results, we plotted the detected CHs using 2CC on the dates from 3 November 2015 through 11 November 2015 (Figure 9). The temporal evolution of the detected CHs close to the solar equator is consistent with the solar rotation. Formation and evolution of a new CH, again close to the solar equator, starting from the 6th of November through 11th of November can also be observed. In addition, temporal evolution of the large CH on the northern solar hemisphere is also consistent in each date (Figure 9).

To further investigate the consistency, we checked the day-to-day temporal evolution of the areas during 2012 and 2016 (Figure 10). Note that the areas are calculated for the last two months of each year. In 2012, there is a general good agreement between our 2CC, CATCH, and HEK CHs especially during December, whereas in November, the HEK CH areas are larger compared to our 2CC and the CATCH (Figure 10a). During 2016, on the other hand, CH areas from the three sources covary with some small differences in amplitudes (Figure 10b).

4. DISCUSSION AND CONCLUSIONS

CHs are the source regions of the steady fast solar winds, which results in CIR driven storms, the so-called HILDCAA events (Tsurutani & Gonzalez 1987). In comparison to their surroundings, CH have lower plasma density and temperatures and therefore they have the lowest emissions in UV and X-ray wavelength range. This physical feature makes them appear as darker regions in passband images of the Sun taken in these wavelengths. CHs are also known to have very complex magnetic structures extending from the photosphere to the corona (Heinemann et al. 2018, 2021), where the open magnetic field lines extend into the interplanetary medium. They also show solar cycle dependence.

There are several methods to identify CHs on the solar images taken by AIA/SDO and EIT/SOHO based on histograms (Krista & Gallagher 2009), multi-thermal intensity segmentation (Garton et al. 2018), and intensity threshold, which is modulated by the intensity gradient of a CH (Heinemann et al. 2019). Recently, unsupervised and supervised ML methods are used to detect CHs using single
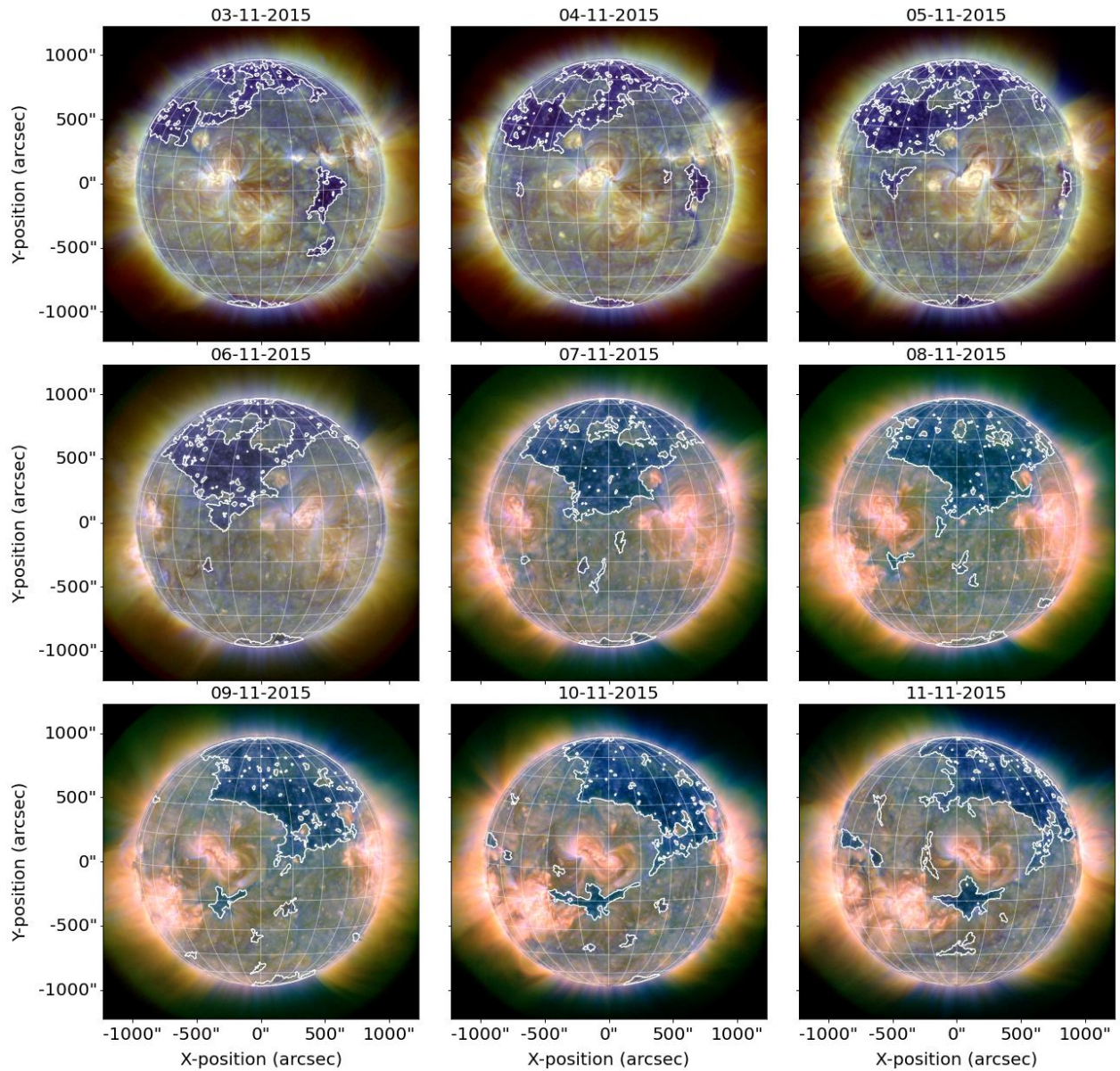
Figure 9. The CH binary maps for a time sequence from 03 through 11 November 2015 identified from the 2CC.

or multi-channel passband data from the AIA/SDO (Verbeeck et al. 2014; Illarionov & Tlatov 2018; Jarolim et al. 2021). The supervised ML methods mainly rely on the CNNs for image segmentation. These methods, however, require a reliable training data set that is CH polygons detected either by an observer or by an unsupervised method.
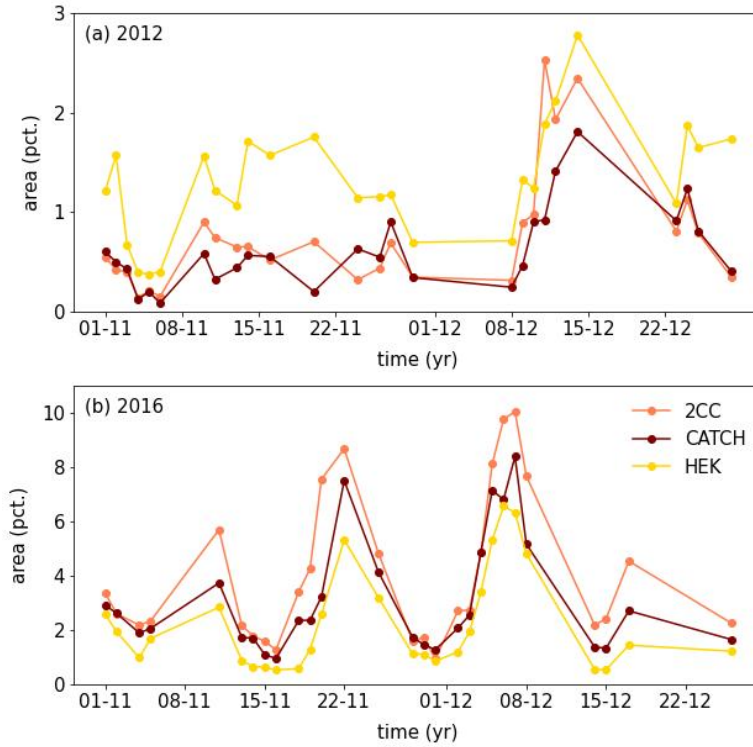
Figure 10. The CH areas during the last two months of 2012 (a) and 2016 (b). The coral, gold, and maroon lines represent 2CC, HEK, and CATCH data, respectively.

In our study, to identify the CHs we used a simple clustering algorithm, k-means, to pixel-wise cluster the passband images of the Sun taken in 171 Å, 193 Å, and 211 Å by the AIA/SDO covering the time period between November 2010 and December 2016. In addition to using a single-channel approach, we used different combinations of these channels. To detect the lower and upper threshold values, we fitted bimodal gaussians to the probability densities of intensities for each channel on each date. We then calculated the thresholds based on the mean and standard deviation of the local maximum at higher intensities. To cluster the passband images, we used the k-means method, where the optimum number of clusters, 3, is calculated based on the scree plot. The k-means method, together with pre- and post-processing steps enabled us to build a automated flexible approach which dynamically responds to day-to-day variations in solar images. As a result we obtained 5 different binary maps for each identified CHs, that are (i) AIA 193, (ii) AIA 211, (iii) 2CC, (iv) 3CC, and (v) 2CO. We then calculated pixel-wise evaluation metrics based on CH binary maps from CATCH and

compared our results with each other as well as those from the HEK database. Following that, we calculated the total percentage area identified as a CH per date, after correcting the binary maps for the projection effects.

Our results show that the 2CC, a composite image using only 193 and 211 Å passband images, provides the best results that is closely followed by results from AIA 193. The median IoU and TSS values for the 2CC are 0.64±0.14 and 0.93±0.06, respectively, while they are 0.62±0.14 and 0.91±0.06 for the AIA 193. Our results show higher similarity to CATCH results than the HEK database (IoU = 0.53±0.13 and TSS = 0.73±0.13). Our results provided better overlap with the CATCH data than those obtained by the CHRONNOS method (Jarolim et al. 2021) for the same period, which provided mean IoU and TSS values as as 0.63 and 0.81, respectively. This method uses all of the 7 channels from the AIA/SDO and line-of-sight magnetograms from the HMI/SDO in a progressively growing CNNs (Jarolim et al. 2021). Even though our results from AIA 193 and 2CC also provide high overall correlations, they are still lower than the correlation coefficient of 0.88 between the HEK binary maps and CATCH.

We also showed the consistency of our results, especially from the 2CC method ,when the formation and temporal evolution of the CHs are considered. Our method was able to identify and track the CHs from the 3 November through 11 November for 9 consecutive days. Additionally, temporal variations of CH areas from our method follows the trends that is observed in the CATCH and HEK CH areas.

To investigate the effects of the chosen lower and upper threshold values, we also calculated the same evaluation metrics and areas for the threshold ranges of $\mu \pm 3\sigma$, $\mu \pm 5\sigma$, as well as for cases where we do not apply any thresholding at all. Similarly, we calculated the thresholds based on the bimodal gaussian fit and the mean and standard deviation of the local maximum at the higher intensities. However, using different thresholds, and also not using any thresholds, provided lower evaluation metrics as well as correlation coefficients of the total areas.

Interestingly enough, our results show significant discrepancies between the identified CHs using our method, HEK and CATCH when we look at the temporal variations in the correlation coefficients

calculated for the total areas. Recently, some steps have been taken to create a reliable database where there is a consensus about the CH boundaries and their uncertainties are being discussed (Linker et al. 2021; Reiss et al. 2021).

In conclusion, as an unsupervised ML method, using the k-means clustering provides better results with those from complex methods, such as CNNs. One of the most important steps in this method is the preprocessing of the data and the choice of the lower and upper threshold values in a more systematic way, which then can lead to automation of the CH detection at any given date or a date range. More importantly, our study shows that there is need for a CH database that a consensus about the CH boundaries are reached by observers independently, and that can be used as the "ground truth", when using a supervised method or just to evaluate the goodness of the models.

ACKNOWLEDGMENTS

REFERENCES

Barnes, W., Cheung, M., Bobra, M., et al. 2020a, aiapy: A Python Package for Analyzing Solar EUV Image Data from AIA, v0.3.1, Zenodo, doi: 10.5281/zenodo.4274931

Barnes, W. T., Cheung, M. C. M., Bobra, M. G., et al. 2020b, Journal of Open Source Software, 5, 2801, doi: 10.21105/joss.02801

Cranmer, S. R. 2009, Living Reviews in Solar Physics, 6, 3, doi: 10.12942/lrsp-2009-3

Delaboudinière, J. P., Artzner, G. E., Brunaud, J., et al. 1995, SoPh, 162, 291, doi: 10.1007/BF00733432

Delouille, V., Hofmeister, S. J., Reiss, M. A., et al. 2018, "Chapter 15 - Coronal Holes Detection Using Supervised Classification, 365–395, doi: 10.1016/B978-0-12-811788-0.00015-9

Eastwood, J. P., Biffis, E., Hapgood, M. A., et al. 2017, Risk Analysis, 37, 206, doi: 10.1111/risa.12765

Garton, T. M., Gallagher, P. T., & Murray, S. A. 2018, Journal of Space Weather and Space Climate, 8, A02, doi: 10.1051/swsc/2017039

Hanssen, A., & Kuipers, W. 1965, On the relationship between the frequency of rain and various meteorological parameters (with reference to the problem of objective forecasting). (Koninklijk Nederlands Meteorologisch Instituut)

Harvey, K. L., & Recely, F. 2002, SoPh, 211, 31, doi: 10.1023/A:1022469023581

Heinemann, S. G., Hofmeister, S. J., Veronig, A. M., & Temmer, M. 2018, ApJ, 863, 29, doi: 10.3847/1538-4357/aad095

Heinemann, S. G., Temmer, M., Heinemann, N., et al. 2019, SoPh, 294, 144, doi: 10.1007/s11207-019-1539-y

Heinemann, S. G., Temmer, M., Hofmeister, S. J., et al. 2021, SoPh, 296, 141, doi: 10.1007/s11207-021-01889-z

Hewins, I. M., Gibson, S. E., Webb, D. F., et al. 2020, SoPh, 295, 161, doi: 10.1007/s11207-020-01731-y

Hong, J., Jiang, Y., Yang, J., et al. 2014, ApJ, 796, 73, doi: 10.1088/0004-637X/796/2/73

Hurlburt, N., Cheung, M., Schrijver, C., et al. 2012, SoPh, 275, 67, doi: 10.1007/s11207-010-9624-2

Illarionov, E. A., & Tlatov, A. G. 2018, MNRAS, 481, 5014, doi: 10.1093/mnras/sty2628

Jaccard, P. 1912, New Phytologist, 11, 37, doi: https: //doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jarolim, R., Veronig, A. M., Hofmeister, S., et al. 2021, A&A, 652, A13, doi: 10.1051/0004-6361/202140640

Karachik, N., Pevtsov, A. A., & Sattarov, I. 2006, ApJ, 642, 562, doi: 10.1086/500820

Krista, L. D., & Gallagher, P. T. 2009, SoPh, 256, 87, doi: 10.1007/s11207-009-9357-2

Lecun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436, doi: 10.1038/nature14539

Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, SoPh, 275, 17, doi: 10.1007/s11207-011-9776-8

Likas, A., Vlassis, N., & J. Verbeek, J. 2003, Pattern Recognition, 36, 451, doi: https: //doi.org/10.1016/S0031-3203(02)00060-2

Linker, J. A., Heinemann, S. G., Temmer, M., et al. 2021, ApJ, 918, 21, doi: 10.3847/1538-4357/ac090a

MacQueen, J. 1967, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (Berkeley, Calif.: University of California Press), 281–297. https: //projecteuclid.org/euclid.bsmsp/1200512992

Marsch, E. 2006, Living Reviews in Solar Physics, 3, 1, doi: 10.12942/lrsp-2006-1

Mumford, S. J., Freij, N., Christe, S., et al. 2021, SunPy, v3.0.3, Zenodo, doi: 10.5281/zenodo.5751998

Paparrizos, J., & Gravano, L. 2015, in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, SIGMOD '15 (New York, NY, USA: Association for Computing Machinery), 1855–1870, doi: 10.1145/2723372.2737793

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3, doi: 10.1007/s11207-011-9841-3

Reiss, M. A., Muglach, K., Möstl, C., et al. 2021, ApJ, 913, 28, doi: 10.3847/1538-4357/abf2c8

Ronneberger, O., Fischer, P., & Brox, T. 2015, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, ed. N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Cham: Springer International Publishing), 234–241

Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, SoPh, 275, 207, doi: 10.1007/s11207-011-9834-2

Schmidhuber, J. 2014, arXiv e-prints, arXiv:1404.7828. https://arxiv.org/abs/1404.7828

Schwabe, H. 1844, Astronomische Nachrichten, 21, 233, doi: 10.1002/asna.18440211505

Schwenn, R. 2006, Living Reviews in Solar Physics, 3, 2, doi: 10.12942/lrsp-2006-2

The SunPy Community, Barnes, W. T., Bobra, M. G., et al. 2020, The Astrophysical Journal, 890, 68, doi: 10.3847/1538-4357/ab4f7a

Tsurutani, B. T., & Gonzalez, W. D. 1987, Planetary and Space Science, 35, 405, doi: https://doi.org/10.1016/0032-0633(87)90097-3

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., et al. 2014, PeerJ, 2, e453, doi: 10.7717/peerj.453

Verbeeck, C., Delouille, V., Mampaey, B., & De Visscher, R. 2014, A&A, 561, A29, doi: 10.1051/0004-6361/201321243

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. 2001, in Icml, Vol. 1, 577–584

Wilcox, J. M. 1968, SSRv, 8, 258, doi: 10.1007/BF00227565

Wyper, P. F., DeVore, C. R., Karpen, J. T., Antiochos, S. K., & Yeates, A. R. 2018, ApJ, 864, 165, doi: 10.3847/1538-4357/aad9f7