SunSCC: Segmenting, Grouping and Classifying Sunspots From Ground-Based Observations Using Deep Learning

Niels Sayez[1], Christophe De Vleeschouwer[1], Véronique Delouille[2], Sabrina Bechet[2], Laure Lefèvre[2]

[1]UCLouvain, B-1348 Louvain-la-Neuve, Belgium
[2]Solar-Terrestrial Centre of Excellence—SIDC, Royal Observatory of Belgium, B-1180 Brussels, Belgium

Key Points:

- We automatically detect, cluster, and classify sunspots using ground-based white light observations and deep learning methods.
- Our method achieves comparable results to those obtained in the literature on space-based continuum and magnetogram observations.
- Ensembles of classifiers are effective in discriminating reliable and likely erroneous predictions on the sunspot classification.

Corresponding author: Niels Sayez, `niels.sayez@uclouvain.be`

Abstract

We propose a fully automated system to detect, aggregate, and classify sunspot groups according to the McIntosh scheme using ground-based white light (WL) observations from the USET facility located at the Royal Observatory of Belgium. The sunspot detection uses a Convolutional Neural Network (CNN), trained from segmentation maps obtained with an unsupervised method based on mathematical morphology and image thresholding. Given the sunspot mask, a mean-shift algorithm is used to aggregate individual sunspots into sunspot groups. This algorithm accounts for the area of each sunspot as well as for prior knowledge regarding the shape of sunspot group. A sunspot group, defined by its bounding box and location on the Sun, is finally fed into a CNN multitask classifier. The latter predicts the three components Z, p, and c in the McIntosh classification scheme. The tasks are organized hierarchically to mimic the dependency of the second (p) and third (c) components on the first (Z). The resulting CNN-based segmentation is more accurate than classical unsupervised methods, with an enhancement up to 16% of F1 score in detection of the smallest sunspots, and it is robust to the presence of clouds. The automated clustering method was able to separate groups with an accuracy of 80%, when compared to hand-made USET sunspot group catalog. The CNN-based sunspot classifier shows comparable performances to methods using continuum as well as magnetogram images recorded by instruments on space mission. We also show that an ensemble of classifiers allows differentiating reliable and potentially incorrect predictions.

Plain Language Summary

Solar activity has long been measured by the number of dark spots, called sunspots, appearing on the Sun's surface and visible from the Earth. Sunspots may be grouped and classified according to a McIntosh classification scheme which describes the sunspot surface structure. This is mostly done manually, which may result in inconsistencies. We use images obtained from the white light telescope of the USET facility at the Royal Observatory of Belgium to automatically detect, group, and classify sunspots. We use Convolutional Neural Networks (CNNs) for the tasks of segmentation and classification, and an algorithm, called 'meanshift' to cluster the sunspots into groups. Even though the USET ground-based images are noisy, and moreover the training data set used for segmentation may be prone to noise, our CNN-based segmentation is more accurate than classical unsupervised methods and is robust to the presence of clouds. When compared to the hand-made USET sunspot group catalog, our automated clustering method was able to separate groups with an accuracy of 80%. Finally, while our sunspot classifier is based only on images recorded by a ground-based instrument, it shows comparable performances to methods using continuum as well as magnetogram images recorded by instruments on space missions.

1 Introduction

Sunspots are dark spots appearing in groups on the surface of the Sun as a manifestation of the solar magnetic activity. The intense magnetic fields embedded within sunspots inhibit convection, cooling the corresponding surface regions. Corresponding areas on the solar surface where temperature has been reduced as compared to their surroundings thus appear as dark spots when viewed in the visible continuum spectrum, also known as White Light (WL). This enhanced magnetic field is the driving force behind the solar variability that influences the space environment of the Earth on a day-to-day basis. Its evolution may lead to magnetic reconnection and subsequent energy release. The stored magnetic energy is then rapidly converted to thermal and kinetic energy, as well as to non-thermal acceleration of particles (Phillips, 1991; Priest & Forbes, 2002). Solar flares, or coronal mass ejections (CMEs) when material is ejected, are understood to be caused by this magnetic reconnection process. Studying sunspot evolution on a long-term basis is also a keystone to several areas of Solar Physics, from helioseismology to solar irradiance modelling, which is directly linked to the evolution of the Earth Climate (Owens, 2013).

The morphology of sunspots is correlated with solar flare occurrence and has therefore received a lot of attention since the end of the 19th century. A classification scheme describing the sunspot magnetic complexity was established by Hale et al., 1919 and is known as the Mount Wilson classification scheme. It groups sunspots into four main classes based on their magnetic structure, that is, on the relative locations and sizes of concentrations of opposite polarity magnetic flux. The first three classes describe the overall mixing of magnetic polarities in sunspot groups: α (unipolar); β (bipolar); and γ (multipolar), whereas the fourth one, known as the $\delta-$configuration, was introduced later on to include close ($< 2°$) mixing of umbral magnetic polarities within one penumbra (Künzel, 1960). Studies have shown that groups that lead to greater magnetic complexity (e.g. $\beta\gamma\delta$ configuration) and larger sunspot area produce flares of greater magnitude at some point in their lifetime (Sammis et al., 2000).

In a similar manner, a classification scheme describing the WL structure of sunspot groups was developed, originally by Cortie, 1901, and later on modified and expanded to include a wider range of parameters (Kiepenheuer, 1953; McIntosh, 1990). Currently this is referred to as the McIntosh scheme, which consists of three components [Zpc]:

- Z: the modified Zurich class, which includes classes from A to H describing the longitudinal extent of the sunspot group. These classes were defined to characterize the process of development and decay of sunspot groups.
- p: the penumbral class, describing the size/symmetry of the largest sunspot's penumbra;
- c: the compactness class, describing the distribution of spots inside a group.

Statistical studies have shown that sunspot groups with higher McIntosh structural complexity classes (corresponding to larger longitudinal extent, large and asymmetric penumbrae, and more internal spots) produced higher flaring rates overall.
The McIntosh scheme is therefore at the basis of several flare forecasting methods which estimate the flare occurrence rate from historical records of flares and sunspot group classes (Bornmann & Shaw, 1994), possibly combining such information with observed waiting time distribution between flares (Gallagher et al., 2002; Bloomfield et al., 2012) or with the evolution of sunspot groups classification (McCloskey et al., 2018).

Attributing a McIntosh class to a sunspot group is thus fundamental to predicting its impact on the Earth space environment. Traditionally, such attribution was done manually, and inserted into catalogs and databases along with metadata related to the number of spots and groups, position, and area of each sunspot. Examples of such catalogs include the merged catalog from Lefevre & Clette, 2014, the US Air Force/SOON catalog (*SOON*, 2008) , Debrecen's Photoheliographic data (DPD), Hungary (Baranyi et al., 2016; Győri et al., 2017), the Spanish observatories catalogs (Aparicio et al., 2022, 2014; Curto et al., 2016), and the Coimbra Observatory catalog, Portugal (Carrasco et al., 2018).

In this work, we use solar observations obtained from the ground-based station USET located at the Royal Observatory of Belgium and we propose SunSCC, a fully automated system addressing three essential tasks regarding the observation of Sunspots on WL images: Segmentation, Clustering of individual sunspots into sunspot groups, and Classification of sunspot groups according to the McIntosh classification scheme (McIntosh, 1990). We focus on the McIntosh classification scheme as it is based solely on WL information, unlike the Mount Wilson scheme, and as it has applications in space weather forecasting.

To our knowledge, this is the first attempt to achieve these tasks using ground-based WL observations only. Most of previous works cited in Section 2 use space-based instruments, and possibly also magnetogram information. As in many other ground-based stations around the world, USET has been observing the Sun for decades without interruption. This allows for a consistent record and statistical studies of sunspot groups over several solar cycles, unlike most of space based instruments which observe during a smaller period of time.

SunSCC works in a cascaded sequence: the output of the segmentation task is given as input to the clustering task, and the clustering is used to locate sunspot groups for which the classification is estimated.

Both the segmentation and the classification use Convolutional Neural Networks (CNN), more precisely the segmentation task uses Unet (Ronneberger et al., 2015), and the classification is based on ResNet (He et al., 2016) followed by Multi Layer Perceptron (MLP, Alpaydin, 2014). Interestingly, our work demonstrates that ensembles of classifiers are effective in differentiating reliable from probably incorrect classifications, based on the consistency of classifier predictions in the ensemble.

For the aggregation of sunspots into groups, we adapt the mean-shift algorithm (Fukunaga & Hostetler, 1975) to the specificities of sunspot groups definition, using the results provided by the segmentation CNN.

We show that, despite the fact that our original data are prone to significant noise due to atmospheric seeing or clouds, and is constituted only of WL observations, our classification method achieves similar performance to methods applied on spaced-based images, possibly using both continuum and magnetogram information. As WL solar observations are extremely common, the SunSCC pipeline could be adopted by many solar observing stations around the globe.

This paper is organized as follows. Section 2 motivates our work and provides a survey of previous related works. Section 3 presents the ground-based observations, and details which specific datasets were used in each task. Section 4 presents the SunSCC pipeline, and the methods used to achieve the three tasks, including the preparation of the necessary training dataset. Quality metrics and a discussion of the experimental results are provided in Section 5. We conclude with future perspectives in Section 6.

2 Motivation and related works

2.1 Sunspot segmentation

Accurate detection of sunspots is fundamental for the construction of solar activity indices such as the International Sunspot Number (Clette et al., 2007), and constitutes the first step towards sunspot classification.

The delineation and classification of sunspots is however partly subjective as two distinct observers might differ in their decisions. Moreover, sunspots are irregular in shape, and have variable intensity and contrast with their surroundings. Ground-based images typically also have different levels of noise and distortion,

In such a non-purely deterministic context, it is particularly useful to rely on the intrinsic statistical robustness of machine learning tools. Automated methods also allow to deal with the ever growing amount of data available and to perform statistical studies on large datasets.

Both region-based and edge-based automated detection methods were used in previous works. Region-based methods encompass on one hand histogram-based segmentation, where pixels are classified according to their intensities (Steinegger et al., 1998

Pettauer & Brandt, 1997 ; Zharkova et al., 2005) possibly combining this with a Bayesian approach (Turmon et al., 2002), and on the other hand region-growing procedures, where the connectivity of individual pixels is used to incorporate information about the local neighborhood (Muraközy, 2022).

Edge-based methods focus on discontinuities and thus on locating region boundaries. Zharkov et al., 2005 used edge detection, followed by mathematical morphology on continuum image from the Michelson Doppler Imager (MDI, Scherrer et al., 1995) space instrument. They compared their results with NOAA synoptic maps produced manually during one year, and found good correlation. Later on, Curto et al., 2008 as well as Watson et al., 2009 used mathematical morphology on WL images, before applying thresholding to isolate the sunspots. Curto et al., 2008 applied his algorithm on WL images from a ground-based station with the goal to reproduce the Wolf number and to compare it against value obtained manually by operators of the same station.
Watson et al., 2009 applied his method on a 1997-2003 MDI continuum dataset to investigate the effect of center to limb variation on sunspot appearance and disappearance.

Carvalho et al., 2020 provided a comparison between intensity-based thresholding and mathematical morphology methods. It showed that mathematical morphology-based approaches perform better than purely intensity-based methods. They also require less pre-processing to remove atmospheric effects in ground-based observations and to correct for limb darkening.

These methods are however still very much dependent on the choice of parameters such as the size of the morphological operators. It is therefore of interest to develop a method which is robust against particular heuristics or parameter's choice. We tackle this problem by considering multiple thresholds per WL image to supervise the training of our machine learning algorithm, as described in Section 4.2.2. The resulting trained neural networks are shown to aggregate and regularize the knowledge extracted from the segmentation masks obtained with multiple thresholds (see Section 5.1).

2.2 Grouping of individual sunspots into sunspot groups

Sunspots are formed in groups that share physical properties such as belonging to the same magnetic flux loops, and are more elongated in longitude than in latitude. Considering a binary mask of sunspots (where '0' means a background pixel and '1' a sunspot foreground pixel), sunspot groups correspond to local maxima in the spatial distribution of foreground pixels inside this binary mask.

Several attempts were made in previous works to find such local maxima. In their work, Curto et al., 2008 used a region growing procedure and predefined criteria of neighborhood to group individual sunspots into a sunspot group. Nguyen et al., 2004 tested various hierarchical algorithms and a simple k-means algorithm.

Since the number of sunspot groups on a given day is not known, the number of clusters k to request from the k-means algorithm can not be defined, making this method impractical. Similarly, finding an appropriate termination condition for the hierarchical algorithms proved to be difficult. The same authors then compared in Nguyen et al., 2006 the performance of several hierarchical algorithms with that of a density-based clustering algorithm known as DBSCAN (introduced in Ester et al., 1996), and found a slightly better performance for the latter.

Finding the local maxima of the statistical distribution that best explains an observed set of samples in a multidimensional space is precisely the problem solved by the so-called 'mean-shift' algorithm (Comaniciu & Meer, 2002). Another reason to choose this algorithm in this work is that it allows for a natural integration of prior knowledge about anisotropic elongation of clusters. This makes it particularly suitable for our purpose since sunspot groups present dominant longitudinal elongation.

2.3 Sunspot Classification

Attributing a McIntosh class to a sunspot group is fundamental for science and space weather purposes. Most of the time, however, such classification is still being done manually by experts. This is a time-consuming process, and although classification rules are well defined, there is room for subjectivity and interpretation, and experts may disagree on sunspot group separation and classification. This in turn results in inconsistencies that stem from these human observation bias as well as non-reproducible catalogs.

To solve the inconsistency problem faced by various solar observatories and space-weather-prediction groups around the world, and to deal with the large amount of solar data available, automated classification of sunspot groups has been envisioned.

Colak & Qahwaji, 2008 combined thresholding and 1-layer artificial neural networks to detect, cluster, and classify sunspot groups according to the McIntosh system. Their work, which represents still a benchmark today, is using both continuum and magnetogram data from the MDI instrument. Abd et al., 2010 used Support Vector Machines on extracted features from MDI continuum images to predict the modified Zurich classification using a one-against all method. They claimed a score of average 80% precision but the method was applied to a restricted dataset (20 samples per class).

The advent of Deep Learning methods allowed computational models that are composed of multiple processing layers to learn representations of data with multiple and hierarchical levels of abstraction (LeCun et al., 2015; Goodfellow et al., 2016). In particular, CNNs have proven superior to previous methods for image classification (Krizhevsky et al., 2012). In sunspot classification, (Knyazeva et al., 2020) used a pre-trained deep neural network and a transfer learning approach to provide an automated McIntosh classification using MDI as well as HMI (Helioseismic and Magnetic Imager, Scherrer et al., 2012) magnetograms. They achieved an average 50% precision score, and significant confusion can be seen between the classes D, E, and F of the modified Zurich class.

Palladino et al., 2022 address sunspot detection and classification using continuum and magnetogram images as well as the Solar Region Summary (SRS) tables of the US Air Force/SOON catalog (*SOON*, 2008) as a source of information about the observed sunspot groups. These tables report the location, composition and McIntosh classification of the visible sunspot groups on a daily basis. A Faster R-CNN model is used to detect sunspots, and Inception v3 for classification. The precision score varies between 30% and 90% amongst classes, and similarly to Knyazeva et al., 2020, confusion between classes D, E, and F of component Z appears clearly.

3 Data

3.1 USET input dataset

Images used in this work were collected by the USET (Uccle Solar Equatorial Table ; *USET*, 1956) station of the Royal Observatory of Belgium (ROB) in the 2002-2019 time period. We used full sun images from the USET White Light (WL) instrument, where images have been pre-processed to center the solar disk in the image at a 1 pixel level accuracy, and the images are translated to have the center in the middle of the image. This dataset is openly available for registered users through the SOLARNET Virtual Observatory (*SOLARNET*, 2021).

The USET station is also equipped with a Visual White light telescope, that projects the image of the Sun on a paper and is used to produce sunspot drawings. Since the 1940s,

Name	Task	Time period(s)	#samples	Annotation type	Annotation origin
$\mathcal{WL}_{SA}$	Seg. (Train)	2013-15	5953	Mask	Automated
$\mathcal{WL}_{SM}$	Seg. (Validation)	2013-15	36	Mask	Manual
$\mathcal{WL}_{C}$	Clustering	2002-12 $\cup$ 2016-19	3150	BBox	SGDB
$\mathcal{WL}_{C}$	Classification	2002-12 $\cup$ 2016-19	5017	McIntosh	DDB, SGDB

Table 1. Summary of the datasets used in this work. BBox stands for 'Bounding Box'. DDB and SGDB are, respectively, the drawing data base and sunspot group database gathered from USET observations.

human operators have made annotations on these drawings regarding the sunspot groups and total sunspot counts. The information contained in these annotations, as well as some information about the drawings have been included in two linked databases: First, the Drawing Database (DDB), which stores the global metadata such as the observer's acronym, total number of spots, sunspot groups, and Wolf number pertaining to one solar drawing, and second the Sunspot Group Database (SGDB), whose entries contain specific information about a sunspot group, such as its heliographic coordinates, area, or manually made classification in the the Zurich and McIntosh systems. Using information contained in the SGDB and the value of the solar radius retrieved from Visual WL observations, it is also possible to determine a Bounding box (BBox) for each sunspot group. Catalogs pertaining to these two DBs have been included in two Table Access Protocol (TAP) services (*USET Database*, 2022). The USET sunspot group catalog stemming from the SGDB was also used in Aparicio et al., 2018, Carrasco et al., 2015, and Lefevre & Clette, 2014.

In this work, the size of the WL images were made uniform to 2048 × 2048 pixels. Indeed, due to a change of camera equipment in June 2008, the image sizes are inconsistent: images recorded earlier than June 2008 are of 1024×1024 pixels size while later ones are of size 2048×2048. To avoid any issue caused by a resolution change, we upsampled earlier images to match the size of later ones using nearest neighbour interpolation.

We detail below which subset will be used for the different tasks of segmenting, clustering, and classifying the sunspots. The characteristics of the various datasets used in this work are summarized in Table 1.

3.2 Segmentation dataset

From the entire dataset, images from the 2013-2015 period were used to train and estimate the CNN-segmentation model. The chosen period corresponds to the peak solar activity during solar cycle 24, offering thus a decent diversity of sunspots and sunspot groups in terms of number and shape.

From the images in this period, a total of 36 images (one image per month) were extracted for segmentation performance validation. These 36 images were annotated manually by ROB experts and constitute the ground-truth for assessing the CNN segmentation's performance. This manually annotated dataset is denoted $\mathcal{WL}_{SM}$. All other images, except the ones captured the same day or the day before or after the one of a test image, form the training set $\mathcal{WL}_{\mathcal{SA}}$, and were processed automatically to generate pseudo-labels to train our CNNs for segmentation, as explained in Section 4.2.1.

Once the CNN-segmentation model has been estimated, it is applied on the WL images within the period 2002-12 $\cup$ 2016-19 (outside the 2013-2015 years) and produces segmentation masks to be used for clustering and classification tasks. The set of WL images recorded during this period, and thus used for both clustering and classification tasks, is denoted $\mathcal{WL}_\mathcal{C}$.

3.3 Clustering dataset

For the clustering, we paired segmentation masks of images in $\mathcal{WL}_\mathcal{C}$ with corresponding sunspot drawings, i.e. with entries in the DDB and SGDB. For each timestamped sunspot drawing, we extracted the chronologically closest WL image in $\mathcal{WL}_\mathcal{C}$ to avoid most of annotation inconsistencies. Such errors may especially occur in the early stages of sunspot group development when classification may change within a few hours.

In our pairings, we set the maximum time difference between drawing annotation and image capture to 4 hours. This selection process produced 3150 pairs of (segmentation masks, sunspot drawings) for the 2002-2012 and 2016-2019 time periods and a total of around 6500 sunspot groups.

3.4 Classification dataset

Input to the classification task are the sunspot groups retrieved by the clustering algorithm. The corresponding number of clusters given to the classification task is of 5017 samples for the period 2002-12 $\cup$ 2016-19. However, sunspot groups that are close to one another deserve a specific treatment, in order to avoid their mixed patterns misleading the classification network during training and validation. Such cases are identified as groups whose bounding box image contains sunspots belonging to other group(s). For each of these groups, before presenting the corresponding image to the classifier, we erase the sunspots of other close group(s) by setting their pixel intensity values to the mean value of the background pixels around their contour. In that way, only one group is available to the classifier at a time.

In the training dataset used in this supervised classification task, each of the 5017 samples gets an associated label that corresponds to the McIntosh class of the matched sunspot group in the SGDB.

4 Methods

4.1 Pipeline overview

The SunSCC pipeline is depicted in Figure 1. It takes as input images obtained from the USET White Light telescope of the Royal Observatory of Belgium and returns as output segmentation masks of individual sunspots, as well as the sunspot groups and their McIntosh classification, with a reliability score. It is divided in three independent blocks as follows.

The first block, dealing with the sunspot segmentation, is addressed by a CNN in the form of Unet architecture (Ronneberger et al., 2015) composed of an auto-encoder with skip connections. The encoding part of the auto-encoder has 5 levels and 4 downsampling steps, and reversely the decoding part has 5 levels and 4 upsampling steps. Each level is composed of two convolutional blocks. The numbers of feature maps (or filters) estimated in each convolutional block at each level are, from highest to deepest level, 32, 64, 128, 256 and 512, respectively.

The second block receives as input the segmentation from the first block, and aggregates the detected sunspots into sunspot groups using an original clustering method inspired by the mean-shift algorithm.

Each identified sunspot group is provided along with an angular distance map to the third block. The latter constitutes a classification network designed to predict the McIntosh 3-components classification. It is composed of a ResNet34 (He et al., 2016) image encoder and three Multi-Layer Perceptrons (MLP) organized in a hierachical way, each having 4 hidden layers and Rectified Linear Units (ReLU) activation functions (Agarap, 2019). Each MLP is specialized in the classification of one component of the McIntosh system. The 4 hidden layers of each MLP contain 512, 256, 256 and 128 neurons, respectively.

An advantage of keeping the three blocks as independent as possible is that each can be used separately as a black-box, modified or enhanced with limited effect on the subsequent blocks. We detail below the methods corresponding to the segmentation, clustering, and classification of sunspot.

4.2 Segmentation

Section 4.2.1 describes our method to build a set of unsupervised but error-prone segmentation masks, using conventional image processing techniques. We then explain in Section 4.2.2 how to use those masks as pseudo-labels to train a CNN model.

4.2.1 Training data preparation, using unsupervised image processing

The generation of unsupervised segmentation masks, also called ‘pseudo-label generation’ in the context of machine learning, is depicted in Figure 2. Each raw image in $\mathcal{WL}_{SA}$ is first processed to correct for limb darkening. Indeed, WL images show a radial decrease in intensity from the center of the disk to the limb, which if not corrected, will hinder threshold-based segmentation methods. To correct this effect, a mask of the limb darkening is created by fitting a polynomial function of the mean intensity of the image in polar projection. The corrected image is then obtained by dividing the raw image by the limb darkening mask.

Next, mathematical morphology operations, in the form of a black top-hat transform is applied, similarly to (Watson et al., 2009). This allows to find dark areas surrounded by a bright environment corresponding to sunspots. Finally, a thresholding is applied to obtain a binary mask.

4.2.2 Training the CNN with stochastic pseudo-labels

The UNet architecture (He et al., 2016) is adopted for our CNN segmentation model since it has proved its accuracy in delineating high resolution segmentation masks in various fields such as medical imaging (Ronneberger et al., 2015; Isensee et al., 2021) and natural images analysis (Siam et al., 2018; Sugirtha & Sridevi, 2022).

Since the smallest sunspots represent only a few pixels and thus may disappear when downsampling large images, we feed the CNN at full original resolution, but use patches of the full image (that we called ‘segment’) of size 512×512 pixels in our segmentation models. Note that unlike during the generation of the segmentation masks by thresholding methods in previous section, the limb darkening effect is not corrected on the images given as input to the CNN. Instead, raw white light images are used. We count on the regularization properties of the CNN to be robust to limb darkening and the presence of clouds.

Finding a single threshold value during the mask generation process is a non-trivial task as the level of illumination may differ not only locally due to the presence of clouds and other artifact source such as atmospheric seeing, but it may also differs from image to image. Moreover, CNNs trained with a systematic labeling error are known to produce predictions with the same error patterns. To circumvent these issues, we consider

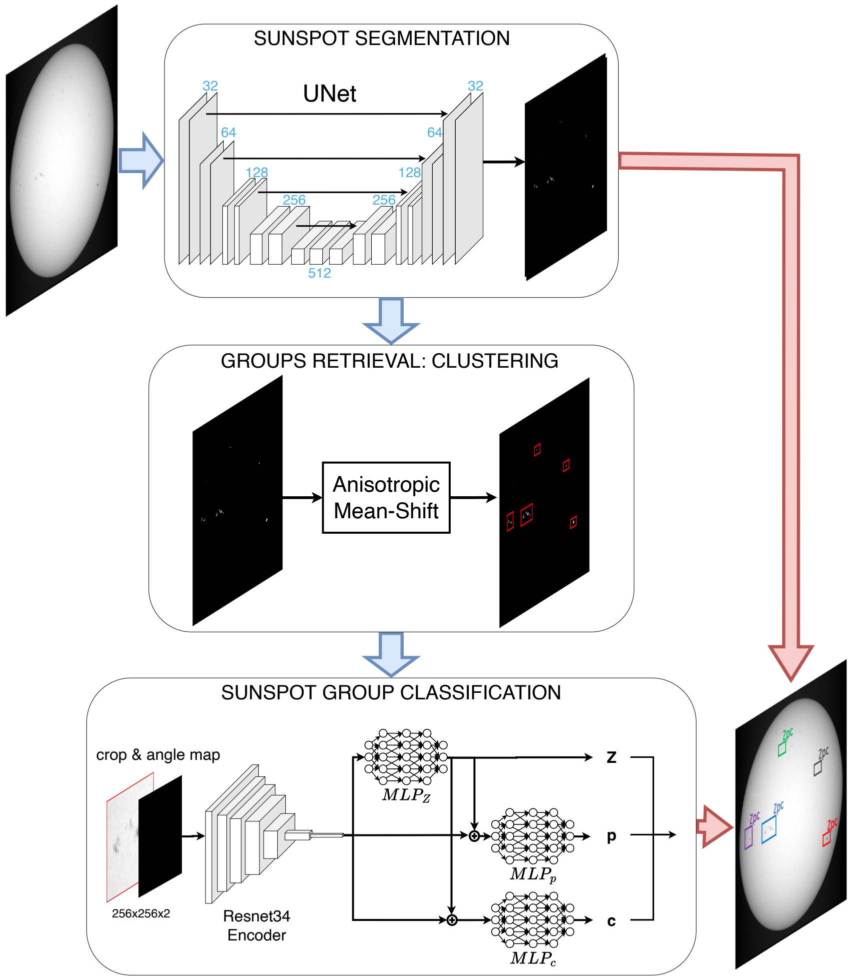


Figure 1. SunSCC pipeline for sunspot segmentation, clustering, and classification. Full-disc images with a resolution of 2048 × 2048 are subdivided into a 4x4 grid, each containing segments of dimensions 512x512. An Unet segmentation network predicts masks for each segment that are reassembled into a 2048x2048 segmentation mask. The Unet network comprises 5 levels (with 32, 64, 128, 256 and 512 filters) and 4 downsampling steps in the encoding part, and reversely 5 levels (with 512, 256, 128, 64, and 32 filters) and 4 upsampling steps in the decoding part. Each level is composed of two convolutional blocks. The detected sunspots are aggregated into sunspot groups by a modified mean-shift algorithm. Each identified group is provided along with an angular distance map to a classification network composed of a Resnet34 image encoder and three MLPs with 4 hidden layers and ReLU activation function. Each MLP is specialized in the classification of one component in the McIntosh system. The 4 hidden layers of each MLP contain 512, 256, 256 and 128 neurons, respectively.

multiple thresholds per WL images, each of them generating a specific mask or pseudo-label, and we train our model using these various pseudo-labels. This is expected to increase the error patterns variability, thereby preventing the CNN to learn them. Our experimental results demonstrate that this original strategy is effective in mitigating the bias introduced by pseudo-labels in the trained model prediction behaviour.

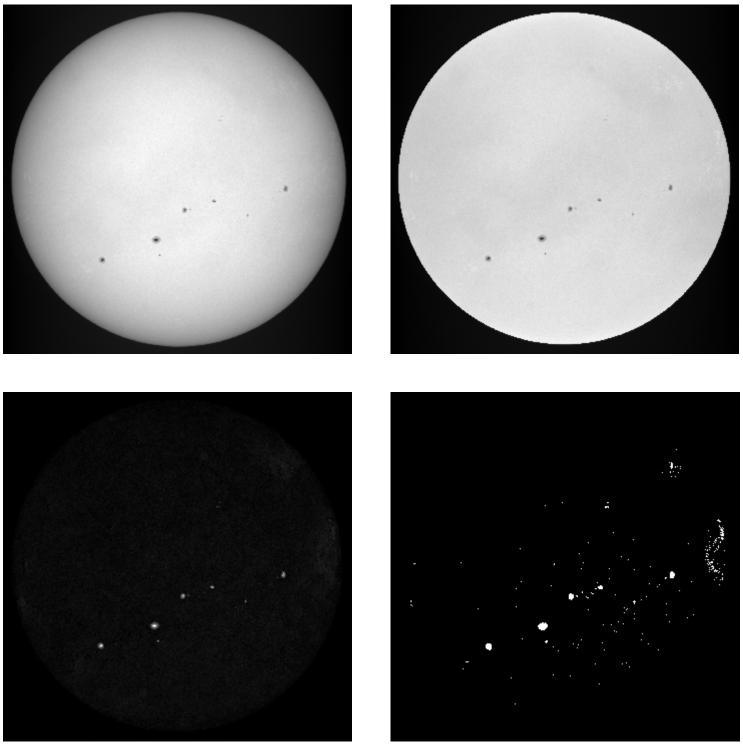


Figure 2. Pseudolabel generation pipeline: (Top Left) Original image, (Top Right) Flattened center-to-limb variation, (Bottom Left) Black-TopHat transform, (Bottom Right) Binary mask obtained with thresholding.

To describe this strategy formally, let I be an image and T a number of thresholds to consider. We denote each threshold by τ_j, with $j \in 1, \ldots, T$. The set of masks (or pseudo-labels) associated to I is denoted

$$\mathcal{M}_I = \{ \; m_{\tau_1}, \; \ldots \; , \; m_{\tau_T} \; \}, \tag{1}$$

where each mask $m_{\tau_j}, j \in 1, \ldots, T$, was created using the image processing method depicted in Figure 2, with a different threshold value τ_j.

When supervising the training of the CNN-segmentation, for each image I, one mask is randomly selected from $\mathcal{M}_I$ in Equation (1) ; where each mask has equal probability to be chosen. This mask is then used to compute the gradients of the binary cross-entropy loss adopted to update the model as is commonly done in segmentation of binary masks (Jadon, 2020).

In our experiments, we used three thresholds ($T = 3$). The central threshold was chosen as the one with highest mask segmentation accuracy on the manually annotated set $\mathcal{WL}_{SM}$. The two other thresholds have been chosen to reject most false positives (at the cost of increasing missed detections) and accept most missed detections (at the cost of an increase of false positives), respectively. The segmentation network was trained using the sum of the cross-entropy loss an the log cosh Dice loss (Jadon, 2020). The AdamW (Loshchilov & Hutter, 2019) optimization algorithm was used with an initial learning rate set to 10^{-4}. The number of epochs was set to 40 with a batch size of 16. The model was trained using the images in $\mathcal{WL}_{SA}$ and their associated pseudo-labels then validated on images in $\mathcal{WL}_{SM}$ and associated manual annotations.

4.3 Clustering

We propose an original method to find sunspot groups in a WL image, given the sunspot masks provided by our CNN segmentation model. Our approach adapts to the context of sunspot group retrieval the mean-shift clustering algorithm introduced by (Fukunaga & Hostetler, 1975) and popularized in computer vision by (Comaniciu & Meer, 2002). The mean-shift clustering algorithm exploits local gradients of a density function to locate its maxima, or modes, given discrete data samples from that function. It simultaneously associates an attraction basin to each of the modes, thereby clustering the observed samples based on the basin they belong to. In the sunspot clustering context, the procedure to discover the modes, and to associate the observed sunspot to those modes, works as follows. Starting from the center of an observed sunspot, it iteratively shifts according to the gradient of the density function obtained by placing at the center of each observed sunspot a kernel shape weighted by the area of the sunspot.

This is repeated until convergence to a local maximum, defining a mode. The convergence procedure is repeated for all the sunspots, which ends up in defining a discrete number of modes, to which sunspots are associated.

The rationale for choosing this particular method is as follows: it adopts a conventional linear kernel, but takes the shape of sunspot group into account, as well as the aspect of individual sunspots when defining the kernel spatial support in longitudinal and latitudinal directions. More precisely, anisotropic kernels are adopted to fit the distribution of sunspots in a group. This originates from prior knowledge that sunspots composing a group tend to be more spread along the longitudinal axis than along the latitudinal axis. Also, the longitudinal bandwidth of each kernel is defined as an increasing function of the area of the corresponding sunspot. The reason why area-dependent kernels are used comes from the observation that bigger sunspots have a larger probability to be a dipole of a large group. Hence, a larger neighborhood should be considered when estimating the density function of the group associated to a bigger sunspot. On the other hand, the kernel bandwidth along the latitudinal axis is kept constant for all the sunspots. Mathematical details of this method are provided in Appendix A.

4.4 Classification

Images from the $\mathcal{WL_C}$ dataset are used to train classification models, with the sunspot group samples obtained by the clustering algorithm described above and tuned with appropriate kernel bandwidth values. As mentioned before, each class is composed of three components: Z, p, and c.

The Z-component, the modified Zurich class, contains classes from A to H. The first three classes represent emerging sunspot groups, from unipolar sunspot (A) to bipolar without penumbrae (B) and bipolar spot with penumbrae on one end of the group (C). Classes D, E, F corresponds to bipolar groups with penumbra on both the leading and trailing sunspot of the group, and are differentiated only by the longitudinal extend of the group: the length between both ends is $\leq 10°$ for class D, between $10°$ and $15°$ for class E and $> 15°$ for class F. Finally, class H corresponds to unipolar sunspot with penumbrae, and usually reflects what remains from a pre-existing bipolar group. Note that classes D, E, and F can easily be differentiated using a post-processing that measures the longitudinal extend between both ends of a sunspot group.

The p-component describes the shape of the penumbrae of the largest sunspot. Class x means no penumbrae, and class r stands for 'rudimentary' penumbrae. Class s and h both represents sunspot groups with mature penumbra of circular or elliptical shape, called 'symmetric'. The north-south diameter across the penumbrae differentiates classes s and h: it is $< 2.5°$ for class s and $\geq 2.5°$ h. Classes a and k represents asymmetric penumbrae, with irregular outline. Similarly to classes s and h, what differentiates a from

k is the length of the north-south diameter across the penumbrae: it should be $< 2.5°$ for s and $\geq 2.5°$ for k, respectively.

Finally, the c-component describes the compactness, that is, how spots are distributed in the space between both ends of a sunspot group. Classes are 'x' (non admissible for unipolar group), 'o' for 'open' (few spots in between two main sunspots), 'i' for 'intermediate' (numerous spots between two main sunspots, but none with a penumbra) and 'c' for 'compact' (area between the two main sunspots is populated with many spots, some of them having a penumbrae).

The McIntosh classification scheme contains as much as 56 admissible classes, and therefore some of them are poorly observed in our training dataset $\mathcal{WL_C}$. On the other hand, some classes have similar morphology and are differentiated only by their sizes. We decided to use a reduced set of classes (Section 4.4.1) to tackle imbalance in class sample size (Section 4.4.2), and to use ensemble classifier to get a measure of confidence in the classification prediction (Section 4.4.3).

4.4.1 McIntosh classes aggregation

We use CNN in order to discriminate the overall aspect of the sunspot groups. In particular, classes that are differentiated only by the sunspot group size or distance between its two ends will be merged together, since a simple post-processing may separate them. This reduction of the set of McIntosh classes operates at three levels.

First the D, E and F classes of the Z-component were merged into a single class denoted LG for 'LargeGroup', since the distinction between these classes is based on the distance between the leading sunspot and trailing sunspot of the group, and not on aspect of any sunspot in the group. Considering the full set of classes might actually lead to confusion of the CNN-classifier between classes D, E and F, as is precisely reported in (Palladino et al., 2022) and (Knyazeva et al., 2020).

Second, regarding the p-component, the symmetric classes, i.e. classes s and h, were merged together in a *sym* class and in an analog manner the asymmetric classes, i.e. classes a and k, were joined in a new *asym* class. Distinction between these classes may easily be done by post-processing, based on a specific length measurement, as explained above.

Third, the number of classes of the c-component was reduced to three by merging classes i and c used for groups where the space between dipoles is populated with smaller sunspots. We named this new class *frag*, standing for 'fragmented'. The distinction between classes i and c is qualitative in nature and hence contains a part of subjectivity. A human operator will have to make the final decision to discriminate them.

The reductions on the usual classification scheme of (McIntosh, 1990) are represented in Figure 3.

4.4.2 CNN-based classification

As shown in Figure 1, our classifier relies on a ResNet34 convolutional backbone to encode the visual information associated to the sunspot group to be classified. The network is fed with standard-sized 256×256 images for all sunspot groups. Before cropping those images, the corresponding full-disk images are rotated to compensate for the sun P-angle, to make the sun rotation axis vertical. The cropped images are composed of two channels: the first channel contains the square segment of the WL image centered at center location of the sunspot group formed during clustering, and the second channel contains the corresponding segment of an angular distance map aligned with the WL image. This second channel is added to the WL input to provide information about the

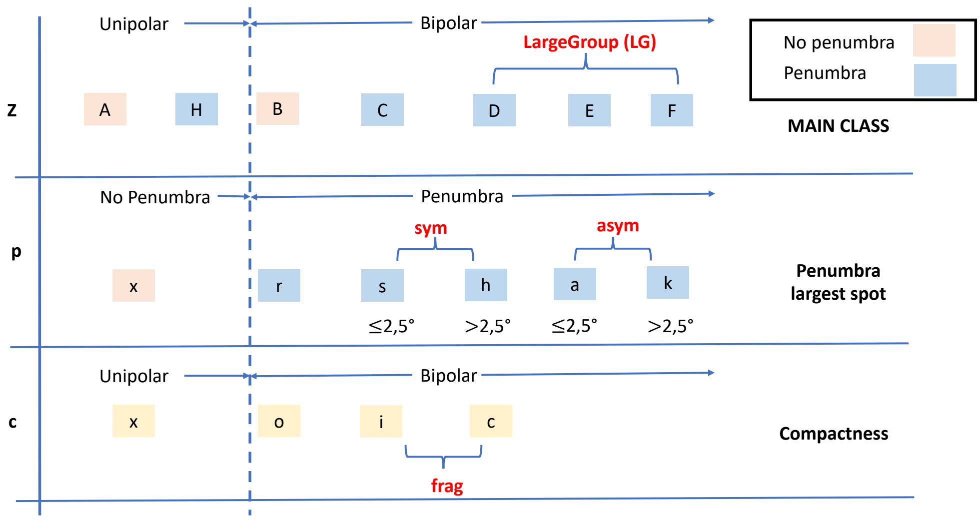


Figure 3. Classification scheme taken from McIntosh (1990). Red fonts indicate how some of the initial McIntosh classes have been merged in a single class in our study (see the text for details).

location of the group on the visible hemisphere to the classification network so that knowledge about distorsion of sunspots shape approaching the limb can be learned.

The embedding representation of the images provided to the ResNet34 backbone is given as input to three MLPs named MLP_Z, MLP_p and MLP_c, respectively specialized in the classification of each of the Z, p and c component of the McIntosh system. The MLPs are organized hierarchically such that the output of MLP_Z is concatenated to the encoded embeddings given as the input of MLP_p and MLP_c. The objective of this architecture is to mimic the dependency of the $p-$ and $c-$component classifications on Z.

Each MLP is trained separately. First the ResNet image encoder is trained along with MLP_Z using a sampling that is balanced in the representation of all the $Z-$classes. Then those two sub-parts (ResNet image encoder and MLP_Z) are frozen and MLP_p is trained using balanced samples of the p-classes. In the final training step, MLP_p is also frozen and MLP_c is trained with a balanced number of sample of each class of the c-component.

In order to balance the representation of the classes for each sub-training, we associate each item i in the training set to three sampling weights, one per McIntosh component. Those weights are inversely proportional to the frequency of the sample class in the training set. These weights control the sampling rate of samples during training.

The classification network was trained using the cross-entropy loss and Adam optimization algorithm (Paszke et al., 2019). The initial learning rate was set to 5.10^{-5}. The batch size was set 16 and the number of epochs was set to 100 for each of the training phases i.e. training of each MLP.

4.4.3 From hard to soft predictions using ensemble of classifiers

In this work, we applied the precepts of *ensemble learning* (Dietterich, 2000; Peake et al., 2020) to soften the prediction of a sample by a single classification network into a confidence measure of belonging to a particular class.

The founding concept of ensemble learning is that the biases of a single classifier are likely to be compensated by biases of other classifiers when combining their predictions.The overall prediction accuracy or ensemble prediction accuracy is expected to be better than the performance of a single classifier (see Dietterich, 2000 for a review). Analyzing the predictions associated to an ensemble of models has also been successfully investigated to quantify predictive uncertainty in neural networks, i.e. to associate a level of confidence to the prediction (Guo et al., 2017; Lakshminarayanan et al., 2017). Cases for which the prediction disparity across models is large are considered to correspond to unreliable predictions. This has been studied and validated in the context of image segmentation (Léger et al., 2022; Dietterich, 2000). In our experiments, multiple instances of the classification network shown in Figure 1 are initialized with different seeds, and are trained independently, to produce multiple classifiers. Hence for each sunspot group, we obtain a pool of predictions. These predictions are then combined into one single output via a majority vote.

The results provided in Section 5.3.4 show that the class prediction inconsistency across the ensemble of classifiers is a valuable clue to identify the cases for which the majority vote prediction is likely wrong.

5 Results and discussion

This section gathers the discussion on the results obtained by the three tasks of the SunSCC pipeline: segmentation (Section 5.1), clustering (Section 5.2), and classification (Section 5.3) of sunspots.

5.1 Sunspot Segmentation

5.1.1 Evaluation metrics

In this work, we evaluated segmentation models as detection models, of which they are a pixel-wise extension. In the performance evaluation of such models, only true positives, false positives and false negatives are considered. On the other hand, true negatives object candidates are left aside as a potentially infinite number of them may lie in each image, making them irrelevant. The performance can thus be measured with the usual *precision* and *recall* that can be combined in a single metric named F1-score. Formally, using the manually annotated dataset $\mathcal{WL}_{SM}$ as ground truth, we first evaluate the quality of the segmentation with the $F1$-score, defined as:

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{2}$$

where TP('True Positive') is the number of correct detections, FN ('False negative) is the number of missed detections and FP('False Positive') is the number of undesired detections. The F1-score thus decreases when there are undetected sunspots (FN) or erroneous sunspot detection (FP). Second, we consider the number of multiple sunspots covered by a single detection (called 'merged' detection) as another quality metric. This is because the precise count of individual sunspots is an important element for the determination of the International Sunspot Number (Mathieu et al., 2019).

Sunspots have a wide range of possible sizes in our images, from a few pixels for the smallest spots to several thousand pixels for the largest spots. Hence, missing a few pixels in the segmentation of a small sunspot has a different meaning than missing the same number of pixels in a large sunspot. Therefore the segmentation of sunspots is analyzed as a function of the sunspot size, by grouping sunspots in five bins with increasing number of pixels. The bin boundaries were chosen so that the population of each bin is sufficient to depict a statistically relevant behaviour in order to reflect the variation of detection performance observed on sunspots of different sizes. The boundaries of these bins are given in Table 2.

bin	0	1	2	3	4
#pixels	< 50	$50-100$	$100-400$	$400-800$	> 800

Table 2. Sunspot size bin boundaries.

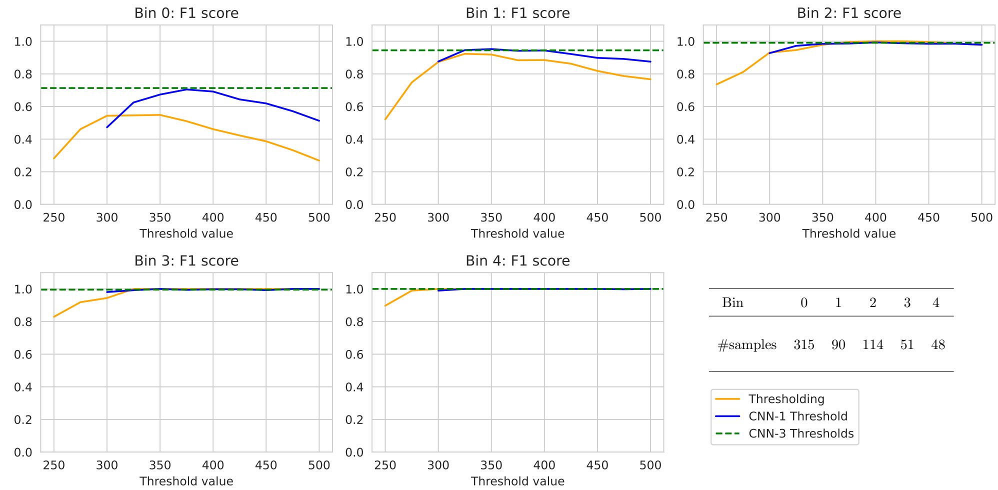
5.1.2 Segmentation results

Figure 4 displays, as a function of the threshold value, the performance of the following three methods: unsupervised threshold-based segmentation, as described in Figure 2, CNN-based segmentation that is trained with one pseudo-label corresponding to one threshold value, and CNN-based segmentation trained with three pseudo-labels per WL images, each pseudo-label being obtained with a different threshold value, and being selected alternatively with equal chance during training. The first two approaches are represented by solid coloured curves, while the third is represented by a constant dashed line as it does not depend on a single pre-defined threshold.

The first noticeable result in Figure 4 is that the CNN-based segmentation regularizes the flaws of small sunspot detection observed when using thresholding. This is observed as the F1 scores of CNN-based segmenters for Bin 0 and Bin 1 consistently outperform those of threshold-based segmenters. Regarding sunspot Bin 0, the highest performance reached using a thresholding method was 55%. The performance increase using CNNs goes from 7% for lower thresholds up to 24% for higher ones, reaching a new maximum of 71%. The same observation could be made on Bin 1 with a smaller increase as the thresholding method was able to achieve a F1-score of 92% F1. Bins of larger sunspots keep similar performance and reach 100% of detections. Details of segmentation performance on the defined bins are shown in Figure 4.

In our experiments, we measured the trade-off between the number of non-detected sunspots and the number of merged sunspots - i.e. fusions - for each threshold-based segmenter and CNN-based segmenter. Measures of this trade-off can be found in Figure 5. One can observe that CNN-based segmentation regularises sunspot detections since the number of misdetections is significantly reduced when compared with thresholding. On the other hand, CNN-based segmentation tends to slightly increase the number of fused sunspots (but not more than a few percent).

Finally, the segmentation using three alternating pseudo-labels per WL image during training achieved performance comparable to that of the best CNN-segmenter using a single pseudo-label, both in terms of F1-score and in terms of fusion versus non-detected sunspots trade-off, as visible on Figures 4 and 5. It also achieves a smaller number of fusion than the thresholding methods with the best F1 scores, as shown on Figure 5. This highlights the benefit of using several pseudo-labels as a solution to train CNN-based sunspot segmentation instead of searching a unique best threshold value. This observation is especially relevant and helpful when few annotated samples are available (in the validation set) to select the unique threshold that would maximize the performance.



Bin	0	1	2	3	4
#samples	315	90	114	51	48

Figure 4. F1-score of sunspot detection as a function of threshold value, for each sunspot bin. The table depicts the population size of each bin. The methods represented are: unsupervised segmentation using for thresholding (orange curve), CNN trained with unique pseudo-label per WL image (blue curve), CNN trained with 3 pseudo-labels per W L image (green dashed line). The last CNN was trained by selecting alternatively with equal chance the pseudo-labels generated using the following threshold values: $\tau_1 = 325$ $\tau_2 = 375$ $\tau_3 = 425$.

5.2 Clustering

To retrieve sunspot groups, we use the masks produced with the best performing CNN classifier in our segmentation experiments. Such masks were obtained for each image in $\mathcal{WL}_\mathcal{C}$ (see Section 3).

The performance of the proposed clustering method is done by measuring the concordance between clusters found by the automatic method, and the SGDB entries. To quantify this concordance, bounding boxes computed from SGDB entries are compared with bounding boxes of groups found by clustering. We refer to SGDB bounding boxes as *target boxes* and bounding boxes of groups formed by clustering as *meanshift boxes*.

The dimension of each target box is computed as a function of the McIntosh class and center position of the sunspot group, and solar radius from the corresponding visual WL observation, see Appendix C for details.

In each image of the set $\mathcal{WL}_\mathcal{C}$, each target box is paired with the closest meanshift box. The set of matchings pairs is denoted M. If two target boxes have the same meanshift box as pair candidate, only the target box with smallest distance is paired, while the other is considered *unmatched*. If both target boxes overlap the meanshift box (and have a zero-distance), the one with largest intersection over union (IoU) is paired. We denote U_T (resp. U_M) the set of unmatched target boxes (resp. the set of unmatched meanshift boxes) along the whole image set $\mathcal{WL}_\mathcal{C}$.

The evaluation of the performance of the proposed clustering method should directly depend on the size of U_T and U_M, denoted $\#U_T$ and $\#U_M$, respectively. Indeed, the unmatched target groups in U_T, and the unmatched mean-shift groups in U_M both correspond to clustering errors, except if they are isolated, i.e. far enough from all groups

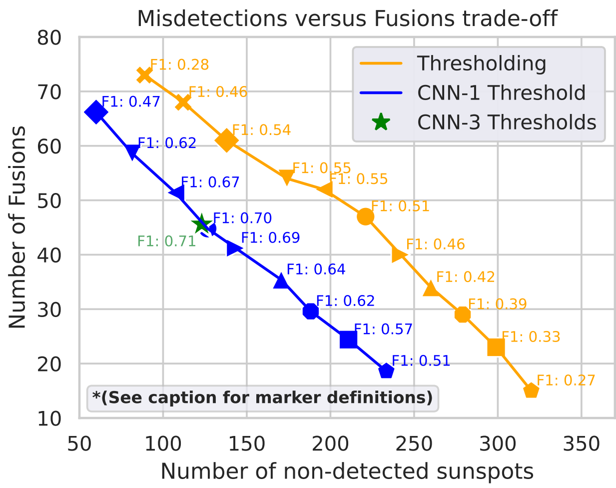


Figure 5. Trade-off between sunspots remaining undetected and sunspots being merged in a single entity. The thresholding method (orange) labels as sunspot the pixels lying above a threshold value in the Black-TopHat transformed image. CNN-1 (blue) corresponds to the models whose training has been supervised with pseudo-labels obtained with a single and fixed threshold. CNN-3 is trained with 3 distinct pseudo-label images per WL image (green). Markers indicate the threshold used (diamond: $\tau = 300$, down triangle: $\tau = 325$, left triangle: $\tau = 350$, circle: $\tau = 375$, right triangle: $\tau = 400$, up triangle: $\tau = 425$, hexagon: $\tau = 450$, square: $\tau = 475$, pentagon: $\tau = 500$, star: $\tau_1 = 325$ $\tau_2 = 375$ $\tau_3 = 425$). The F1 scores indicate performance on sunspots in the first bin, which corresponds to the smallest and thus most challenging sunspots to detect.

of the other [target/mean-shift] category. Indeed, when the sunspot segmentation is accurate, U_M corresponds to the cases for which a target group has been erroneously split into multiple meanshift groups (leading to unmatched meanshift groups), and U_T to the case for which a meanshift group has aggregated two target groups (leading to an unmatched target). When the segmentation is not correct, false positives and missed detections lead to isolated mean-shift and isolated target groups, respectively. We let $I_T \subset U_T$ and $I_M \subset U_M$ denote the unmatched bounding boxes associated to those missed or false positive sunspot detections. Since they reflect detection errors, those I_T and I_M bounding boxes should preferably be ignored when assessing the clustering performance in order for the evaluation of the clustering method to be as independent as possible from the quality of the segmentation. Hence, $\#I_T$ and $\#I_M$ are removed from $\#U_T$ and $\#U_M$, respectively, when computing the clustering accuracy metric.

Formally, the accuracy of the clustering algorithm is computed using the following formula:

$$\text{Accuracy} = \frac{\#M}{\#M + (\#U_T - \#I_T) + (\#U_M - \#I_M)} \tag{3}$$

Appendix B presents a grid search to find appropriate values for the latitude bandwidth as well as maximum longitude bandwidth to define the kernels used for clustering. This reveals that our method achieves 80% accuracy when appropriate meanshift parameters are adopted. The values of longitudinal and latitudinal kernel bandwidths achieving this performance are .4 radians and .1 radians, respectively, see Figure B1 in Appendix B. These values have been adopted when creating the dataset on which the classification network was trained.

5.3 Classification

The details about the distribution of samples among the classes of the Z, p and c components of the McIntosh system are shown in Tables 3, 4 and 5, respectively. In our experiments, we used 85% of the available samples for training and 15% for testing.

One can observe the presence of class imbalance in each of those tables as the representation of rarest classes can vary from half to the third of the representation of the most populated ones. Also, the respective imbalances of the classes in Z, p and c components make it difficult to train the whole network at once as balancing the training on one component does not guarantee balanced training for classes of other components. The context of this work thus requires a training supervision strategy of the classification models that takes this additional complexity into account to ensure balanced representation of each class for each McIntosh component. As explained in Section 4.4.2, we use weights associated to each item to control the sampling rate during training.

	A	B	C	LG	H
#samples	578	691	1026	1377	1345

Table 3. Class distribution of Z component in classification dataset.

	x	r	sym	asym
#samples	1269	533	1845	1370

Table 4. Class distribution of p component in classification dataset.

	x	o	frag
#samples	1923	2014	1080

Table 5. Class distribution of c component in classification dataset.

5.3.1 Mean classification accuracy as a function of the category

In our experiments, we trained 20 classification CNNs to produce ensemble predictions. Each test sample was provided to every CNN such that a majority vote could be performed on the predictions of each component Z, p and c. The performance of this ensemble learning approach was measured with the weighted F1-score on each component, and this for two sets of test samples: first, the set of sunspot group samples that lie far away from any other group on the cut-out WL image given to the CNN, and second the set of samples close to another group, either because their Bbox overlaps that of other group(s) and/or because other group(s) appear on the WL image given to the CNN. On the first set, the proposed network achieved 69% F1-score on the classification of the Z component, 62% on p and 72% on c. These results solely based on ground-based images are comparable with the solution proposed by (Palladino et al., 2022) that uses magnetograms along with satellite based WL image. Their solution respectively reached 67% F1 score for the Z component, 55% for p component and 73% for c component. On the second set of harder samples, the proposed network achieved 57% F1-score on the classification of Z component, 54% on p and 63% on c. Lower performances were expected as it is more likely that these samples have suffered errors during clustering and therefore do not present the characteristics of the McIntosh class they are labeled with. One can expect that further enhancement of the clustering method allows to reach classification performances on the second set of samples similar to the first, provided that the unwanted sunspots in the input image are hidden. The results presented from now on depict the performance on the set of sunspot groups that are far from any other group.

	Z	p	c
Isolated groups	69%	62%	72%
Packed groups	57%	54%	63%

Table 6. F1 score of the classification with majority vote for Z, p and c components on test sunspot groups that are far from other group(s) (top) or close to, i.e. with overlapping bounding boxes, other group(s) (bottom).

The breakdown of our results are shown in Table 6, Figure 6 as well as Tables E1, E2 and E3 from Appendix E. Beside the F1-score, these tables also present the Precision and Recall metrics, which are defined in Appendix D. Figure 6 shows confusion matrices presenting the performances of the classification of the Z, p and c components. Formally, each entry $\mathbf{M}(i, j)$ of the confusion matrix $\mathbf{M}$ denotes the rate at which the ground-truth class i is assigned to class j. On these matrices, one can observe that the majority votes lead to a significant amount of correct predictions. Indeed, the diagonal entries, which represent the per-class recall, have high values, except for the entry corresponding to class C. This exception is expected due to the intermediate status of class C between class B and the class grouped under LG (McIntosh, 1990).

The next three sections presents three take-away messages that can be retrieved from our experiment: performance of CNN-classifier decreases near the limb, even if some information about limb is provided to the algorithm, the CNN-classifier is able to find annotation errors in the SGDB, and finally the ensemble classifier allows to identify ambiguous cases in sunspot classification.

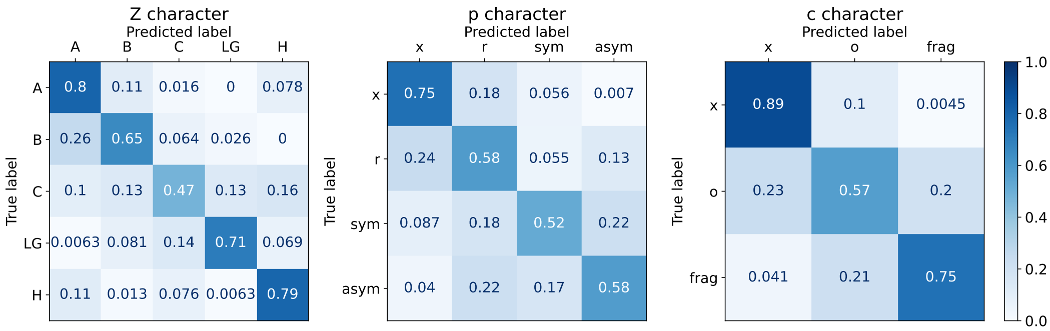


Figure 6. Test set confusion matrices of classification with majority vote for the classification of the Z (left), p (center), c (right) components. In each matrix, the rows represent the true label of test samples while columns represent the label predicted by majority vote. The matrices are row-normalized such that per-class recall values are shown on the diagonals. See Appendix E for details about recall metric computation.

5.3.2 Classification accuracy as a function of the angular distance to the solar disc center

Figure 7 (top-left) depicts the classification accuracy of each Z, p and c component as a function of the angular distance to the Sun center. This distance between a sunspot group and the Sun center is computed as $\arcsin(r/R)$ where r is the distance between the group centre and the solar disc center and R is the solar radius. We observe that the mean accuracy changes with the angular distance. This variation is not due to the stochastic nature of the training procedure (Alpaydin, 2014), as shown by the small error bars presented on Figure 7 (top-left) to depict the standard deviation resulting from multiple stochastic trainings of the neural network. Instead, a careful analysis of the distribution of the classes in the test and training sets of each component has revealed that the drops of mean classification accuracies observed for some angular distances are primarily due to a lack of training samples for some classes that are significantly represented in the test set. In other words, they are caused by a mismatch between the distribution of classes in the test and the training sets when the number of training sample is small. Hence, building the training set so that all classes are reasonably represented across all angular distances is expected to mitigate the erratic fluctuations of accuracy observed in the plot.

This interpretation is supported by Figure 7 (bottom), which presents, for the 70−80 degrees bin, the distributions of the labels associated to the p-component in the training and test sets. We observe that the 'sym' class is dominant in the test set, but only poorly represented in the training set (less than 20 samples). This explains why the corresponding classification accuracy sharply drops in the graph presented in the top of the figure. It is worth noting that this mismatch between the distribution of classes in the test and the training sets is especially problematic when the number of training sample is small, since it ends up in allocating a small number of training samples to a class that is largely represented in the test set, thereby hampering the overall test accuracy.

Note also that the variable number of samples available around a given angular distance is directly linked to the distribution of sunspots on the Sun over time -e.g. the butterfly diagram (Hathaway, 2015)-, combined with the position angle between the geocentric north pole and the solar rotational north pole (P). As P varies over the course of a year, the solar equator moves around the solar disk center creating a non-uniform distribution. This effect is depicted in Figure 7 (top-right), which presents the training sample distribution as a function of the angular distance. It peaks between 20 and 30 degrees, and reveals the small number of samples available in the bins with small or large angular distance.

A few examples of failure cases near the limb are shown in Figure 8. The example on the left (August 9, 2005, 11:22) shows a group on the limb that is recorded as H-asym-x in the SGDB and that is classified as C-r-o by the network. On the drawing, one can see that the observer probably classified it as A-x-x with 2 spots although a penumbra was drawn. The operator encoding the group in the catalog saw the penumbra and classified it as H. However, considering that there is a non-negligible distance between both spots and that this distance is further underestimated due to projection effects at the limb, it could also be considered as C-class. The fault in such failure cases is not to be blamed on the classification network only, nor on the operator. This illustrates complications that may occur due to the position of the group close to the limb.

5.3.3 Finding annotation errors

Discrepancies may exist between the classification recorded in the SGDB and what can be seen on the temporally closest WL image. There are multiple explanations for this. (1) What an observer sees when annotating a drawing and what the digital camera captures are different, e.g. due to turbulence. Most of the time, discrepancies would be in the other direction, i.e. the observer has a long integration time when observing spots and groups while the WL image has routinely limited integration capabilities. (2) A human mistake may occur when saving new entries in the SGDB. An example of such discrepancies that occurred on 6/09/2004 is shown in Figure 9. On this figure one can see that the observer wrote "C" but did not draw the penumbra correctly. Therefore, the operator wrote "B" in the USET catalog as the drawing is historically the unique

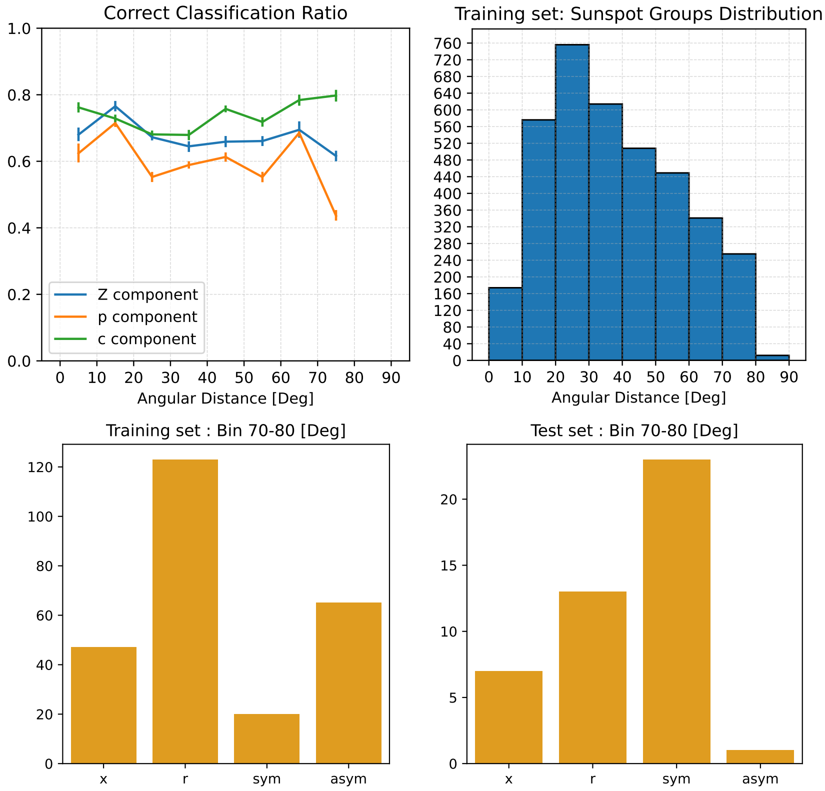


Figure 7. (Top Left) Classification accuracy along angular distance with error bars shown as vertical lines. (Top Right) Distribution of test set samples along angular distance. 0 is center of the solar disk, 90 is the limb. (Bottom Left) Distribution of the train samples of bin 70-80 degrees among classes of the p component.(Bottom Right) Distribution of the test samples of bin 70-80 degrees among classes of the p component.

source to encode groups. However it was C-class as clearly visible on the corresponding WL image. Still, the classification network was able to predict the classification correctly, but it was counted as an error in the performance evaluation due to this erroneous entry in the SGBD.

5.3.4 Identifying ambiguous cases

As stated above, using an ensemble learning approach allows to put a confidence measure on the overall classification of a sample. A noticeable result of our experiments is that correct ensemble predictions have low discrepancies between the votes while incorrect ensemble predictions have their votes more distributed among the classes. When the majority of predictions voted for the correct class, we observe that, for each McIntosh component, more than 93% of the predictions participated to the correct majority vote, as can be seen in top row matrices of Figure 10. On the other hand, only a relatively low percentage of predictions that participated to an incorrect majority vote did vote for the correct class, as can be seen by the diagonal values in the confusion matrices of bottom row matrices of Figure 10. This behavior allows for the identification of difficult cases that may be submitted to an expert for correct classification when clas-

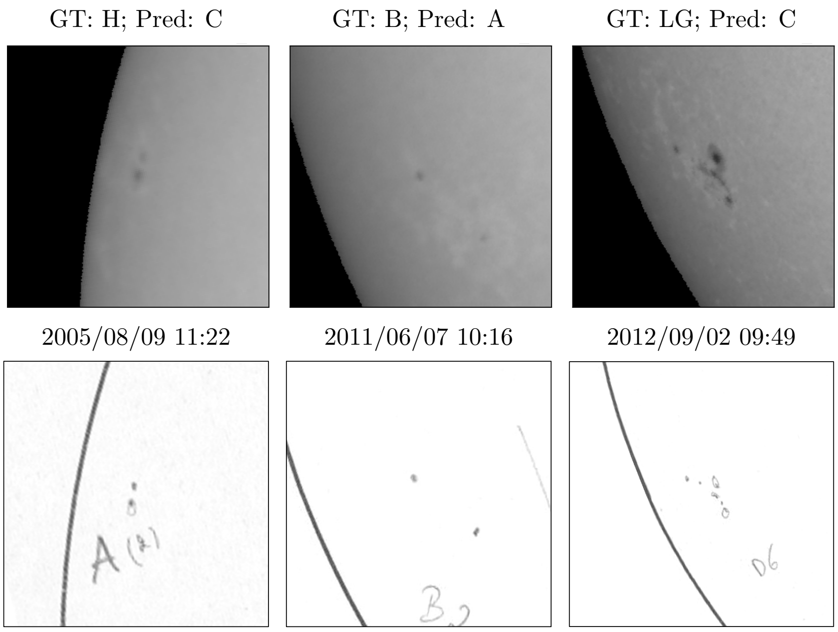


Figure 8. Examples of failure cases close to the limb. For each sample, the capture date and time as well as the expected ground truth classification (GT) and the predicted classification (Pred) are given, along with the WL image presented to the classifiers (top row) and corresponding drawing with annotations about sunspot count and class written during observation (bottom row).

sifying new groups and can also help to raise a flag to signal the presence of potential annotation errors when trying to classify sunspot groups from the past. Examples of erroneous low-confidence majority vote predictions are shown in Figure 11.

6 Conclusion and future prospects

We proposed SunSCC, a comprehensive pipeline that segments sunspots observed in ground-based white light images using a convolutional neural network, aggregates them into groups by an original clustering method, and provides their 3-components McIntosh classification with a second convolutional neural network. Our experiments demonstrate that our CNN-based segmentation is able to retain the best out of pseudo-labels, to predict more accurate masks than a conventional unsupervised image processing approach. Also, alternating between multiple pseudo-labels for each sample during CNN training results in a model with performances similar to the ones obtained when training with a single but carefully selected (i.e. using manually annotated data) pseudo-label. The sunspot aggregation method we proposed is able to retrieve 80% of the sunspot groups surveyed by the USET facility of the Royal Observatory of Belgium (ROB) during the 2002-2019 time period. Finally, our sunspot group classification neural networks achieved performances on par with existing solutions using satellite-based images combined with magnetograms, while relying solely on ground-based images.

In our experiment to find the best CNN-classifier, we tested the introduction of masks that would separate the umbrae and penumbrae in a sunspot using a method based on canny edge detection. This did not improve the results of the CNN-classifier, and hence

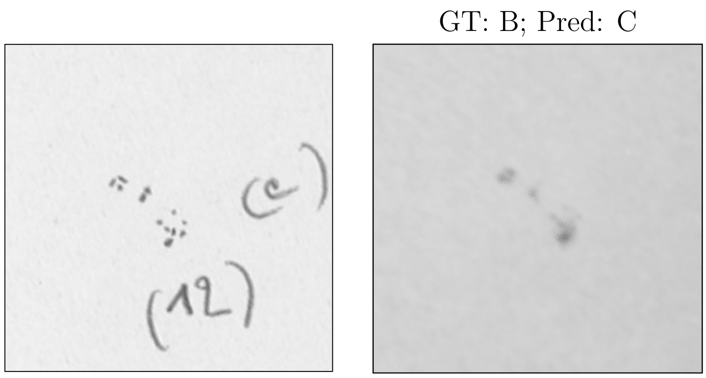


Figure 9. Side by side comparison of drawing annotation (Left) and corresponding WL image given as input for CNN-based classification (Right) for sunspot group observed on September 6, 2004. The observer annotated this group as C-class without penumbra delineation. When encoding in the SGDB, this group was classified as B-class for this reason. The classification network was able to retrieve the correct class.

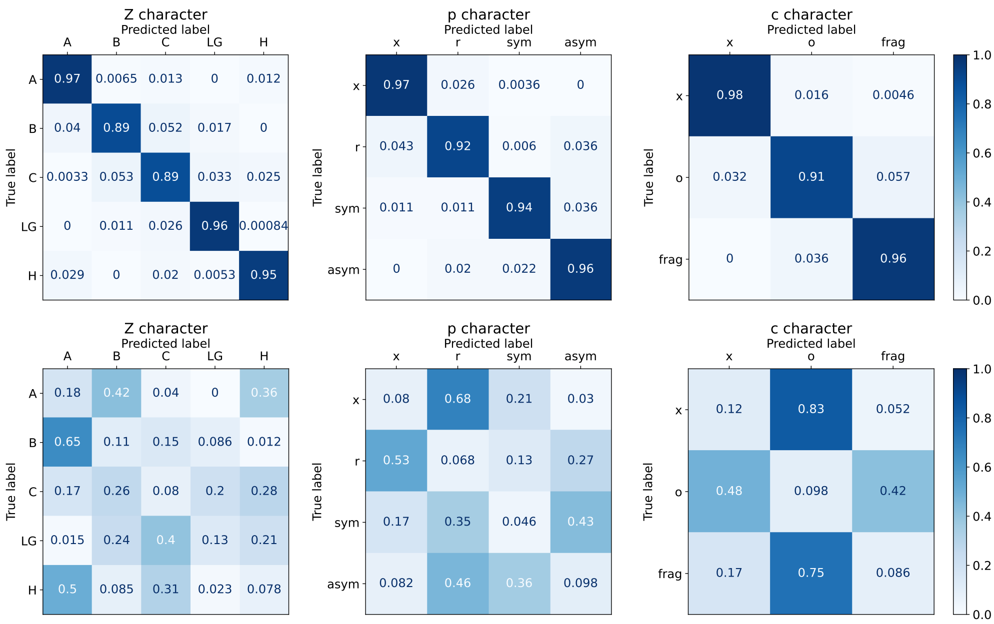


Figure 10. Per-component test set confusion matrices of predictions that participated to a correct majority vote (top row matrices) or incorrect majority vote (bottom row matrices). In each matrix, a cell shows the rate of predictions that voted for its column class while the row class was expected. The more intense the blue, the higher the rate. In the top row matrices, off-diagonal values represent erroneous votes that might occur, even in case of a correct majority vote. In the bottom row matrices, some of the predictions voted for the correct class, hence the non-zero values along the matrix diagonal.

was not presented here. In the near future, we plan however to produce these masks of

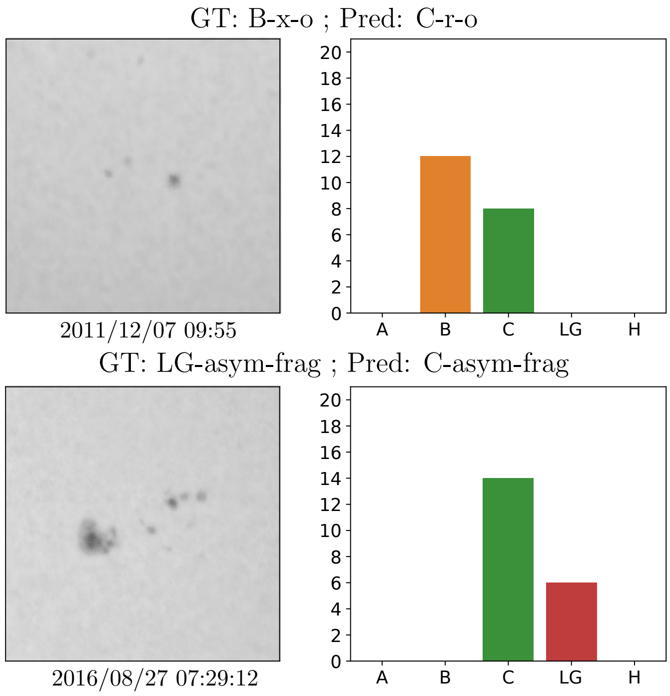


Figure 11. Examples of ambiguous cases. For each sample, the expected ground truth (GT) classification and the predicted class (Pred) are shown as well as the distribution of votes taken into account in the majority prediction of the ensemble of classifiers.

umbrae and penumbrae, and compute relevant statistics such as areas, so that they can be automatically included in the USET Sunspot Group Database. We also plan to include in SunSCC a module for sunspot tracking, and to add this information in the SGDB, to help future research e.g. on solar cycle studies.

On the operational level, we plan to have the SunSCC pipeline running at ROB, and producing daily sunspot masks, grouping, and classification. This will be an aid to human operator, and we will be able to operate quality control on the SGDB. Experts will not be replaced completely by the pipeline as the classification accuracy of the deep learning models is currently not sufficient for complete autonomy and still requires human post processing to comply with the whole spectrum of the McIntosh classes. Instead, it may be used as an assistant tool that can draw attention on certain cases and provide automated preliminary classification that may be overturned by experts if needed. Having an automated McIntosh classification will also be helpful for flare forecasting in ROB's space weather operational services. Moreover, SunSCC could potentially be installed at any ground-based station equipped with a white light telescope. A Transfer Learning approach could indeed be used to adapt the CNN-segmentation and classification to other ground-based instruments.

Open Research

Software and Data availability statement

The image data used in this work is available through the SOLARNET Virtual Observatory `https://solarnet.oma.be/` (*SOLARNET*, 2021). The information from the

Drawing data base and Sunspot Group Database is available from the 'USET drawings' and 'USET Groups' Table Access Protocol service, accessible e.g. through `https://vo-tap.oma.be/` (*USET Database*, 2022). The code related to this work is publicly available in an archive on the FAIR compliant platform Zenodo at `https://doi.org/10.5281/zenodo.8236706` (Sayez, 2023). The archive contains code to generate datasets for each task as well as code for CNN training and performance evaluation.

Acknowledgments

The authors gratefully acknowledges funding from the Belgian Federal Science Policy Office (BELSPO) through the BRAIN DeepSun project (B2 /191/P2/DeepSun). Part of the work was sponsored by the Solar-Terrestrial Centre of Excellence, a collaboration between ROB, the Royal Meteorological Institute, and the Royal Belgian Institute for Space Aeronomy funded by BELSPO.

Appendix A Clustering method details

Here we provide details on the clustering method. Let S_I be the set of sunspots in an image I, each sunspot $s \in S_I$ is defined by its latitude coordinate l_s, its longitude coordinate L_s and its area a_s in *millionth of hemisphere* [μHem]. The probability density function associated to the observed sunspots in an image I is defined as a weighted sum of kernels

$$pdf(l, L) = \sum_{s \in S_I} a_s * k\left(\left(\frac{l - l_s}{h}\right)^2 + \left(\frac{L - L_s}{w(a_s)}\right)^2\right) \tag{A1}$$

with h the kernel bandwidth along the latitude axis, $w(a_s)$ the bandwidth along the longitude axis and $k(\cdot)$ the linear kernel function defined as

$$k(d) = \begin{cases} 1 - d, & \text{if } ||d|| \leq 1. \\ 0, & \text{otherwise.} \end{cases} \tag{A2}$$

The value of $w(a_s)$ is defined in the $[h, w_{\max}]$ interval according to the following formula:

$$w(a_s) = h + \alpha_s(w_{\max} - h), \tag{A3}$$

where α_s is the longitudinal expansion factor, defined as

$$\alpha_s = \min(\frac{a_s}{a_{\max}}, 1). \tag{A4}$$

and $w_{\max}$ is a parameter defining the maximal kernel bandwidth along the longitude axis to associate with sunspots with area over a_{max}. In our experiments, the parameters h and $w_{\max}$ were set to $h = .1$ [rad] and $w_{\max} = .4$ [rad] (see Appendix B for details). We set the value of $a_{\max}$ to 200 [μHem], which is close to the 75th percentile value of the area of a C class sunspot group as surveyed by the ROB.

As a result, during the shifting step, the latitude coordinate is updated according to the following formula

$$l \leftarrow \frac{\sum\limits_{s \in N(l,L)} a_s * l_s}{\sum\limits_{s \in N(l,L)} a_s} \tag{A5}$$

where $N(l, L)$ is the set of samples lying in the elliptic neighborhood where the kernel derivative is non zero. Given Eq.(A2), this corresponds to locations (l_s, L_s) for which $\left(\frac{l - l_s}{h}\right)^2 + \left(\frac{L - L_s}{w(a_s)}\right)^2 \leq 1$, where the derivative of the kernel function is equal to 1.

The longitude coordinate is updated in an analogous fashion.

This process converges in only a few iterations.

Appendix B Optimal Clustering parameters grid search

The accuracy of the proposed clustering method using each combination of latitudinal kernel bandwidth and maximal longitude bandwidth is shown on Figure B1. We considered the [.02, .12] radian range for the latitudinal bandwidth with a step of .02 and the [.1, .75] radian range with a step of .5 for maximal longitudinal bandwidth. For each combination, we applied clustering on images in $\mathcal{WL}_C$ and measured the accuracy of the sunspot group recovery. Our method achieves 80% accuracy when the latidudinal and longitudinal kernel bandwidths are set to .1 radians and .4 radians, respectively.

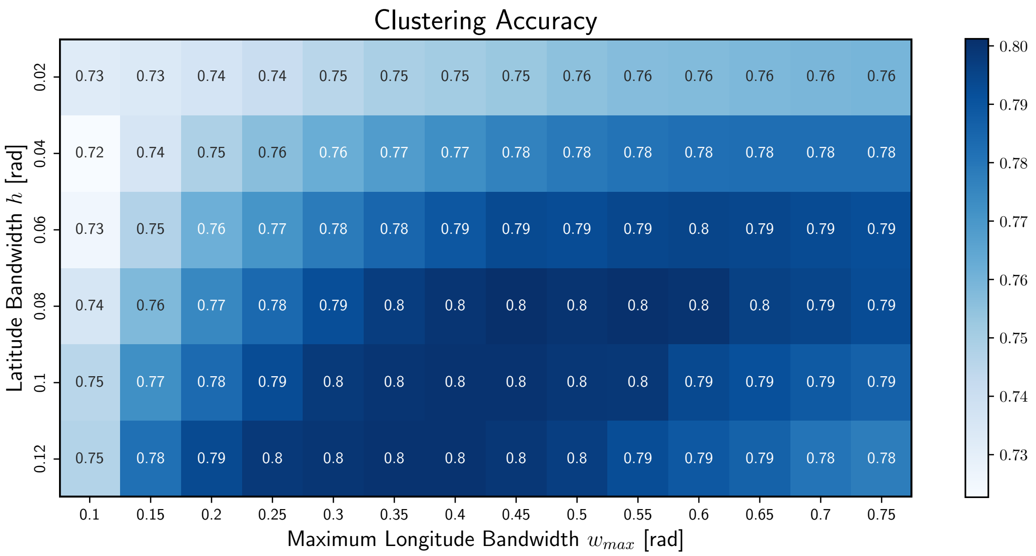


Figure B1. Search for optimal values kernel longitudinal and latitudinal bandwidths for the proposed clustering method applied on images in $\mathcal{WL}_C$. Each cell shows the accuracy of the clustering tuned with kernel dimensions of the corresponding row and column.

Appendix C SGDB entry bounding box computation

The dimension of each target box is computed as a function of the McIntosh class, latitude l and longitude L of the sunspot group center, and solar radius R from the corresponding visual WL observation. Target boxes dimensions along longitudinal axis B_L and latitudinal axis B_l are computed as:

$$B_L = s_Z \cdot |\ \cos(\ L\ )\ | \quad ; \quad B_l = s_Z \cdot |\ \cos(\ l\ )\ | \tag{C1}$$

where s_Z is the length of a side of the sunspot group squared bounding box as a function of the class of the Z component. The values of s_Z at Sun center are given in Table C1 for each class of the $Z-$component in the McIntosh classification system. These values were obtained by expanding the typical size of each Zurich class described in (McIntosh, 1990) to take into account the tilt and P-angle and are expressed as a function of the radius of the solar disc in the processed image.

	A	B	C	D	E	F	H
Box side	$R/30$	$R/9$	$R/9$	$R/9$	$R/6$	$R/4$	$R/18$

Table C1. Length of one side of a squared bounding box for sunspot group located at Sun center and following a particular class according to Z-component classification in the McIntosh system. R denotes the sun radius on the corresponding visual WL observation.

Appendix D Classification precision, recall, support and F1 score

The precision, recall and F1 score metrics for classification are computed using the following formulae

$$precision = \frac{TP}{TP + FP} \tag{D1}$$

$$recall = \frac{TP}{TP + FN} \tag{D2}$$

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 * precision * recall}{precision + recall} \tag{D3}$$

where TP are true positives, FP are false positives and FN are false negatives.

The weighted average F1 score for classification with multiple classes is calculated by taking the mean of all per-class F1 scores while considering each class's support i.e. the number of occurrences of the class in the set.

Appendix E Classification results details

Class	Precision	Recall	F1-score	Support
A	0.5	0.8	0.61	64
B	0.57	0.65	0.61	78
C	0.59	0.47	0.53	120
LG	0.86	0.71	0.77	160
H	0.78	0.79	0.79	158
Weighted Avg	0.7	0.68	0.69	580

Table E1. Ensemble approach classification performance on classification of Z-component. See Appendix D for details about computation of columns.

Class	Precision	Recall	F1-score	Support
x	0.74	0.75	0.75	142
r	0.24	0.58	0.34	55
sym	0.73	0.52	0.61	208
asym	0.66	0.58	0.61	175
Weighted Avg	0.66	0.6	0.62	580

Table E2. Ensemble approach classification performance on p-component. See Appendix D for details about computation of columns.

Class	Precision	Recall	F1-score	Support
x	0.77	0.89	0.82	222
o	0.73	0.57	0.64	235
frag	0.66	0.75	0.7	123
Weighted Avg	0.73	0.73	0.72	580

Table E3. Ensemble approach classification performance on c-component. See Appendix D for details about computation of columns.

References

Abd, M. A., Majed, S. F., & Zharkova, V. (2010). Automated classification of sunspot groups with support vector machines. In K. Elleithy, T. Sobh, M. Iskander, V. Kapila, M. A. Karim, & A. Mahmood (Eds.), *Technological developments in networking, education and automation* (pp. 321–325). Dordrecht: Springer Netherlands.

Agarap, A. F. (2019). *Deep learning using rectified linear units (relu).*

Alpaydin, E. (2014). Multilayer perceptrons. In *Introduction to machine learning* (p. 267-316).

Aparicio, A. J. P., Carrasco, V. M. S., Gallego, M. C., & Vaquero, J. M. (2022, May). A Sunspot Catalog by Rafael Carrasco at the Madrid Astronomical Observatory for the Period 1931 - 1933. *Solar Physics*, *297*(5), 58. doi: 10.1007/s11207-022-01992-9

Aparicio, A. J. P., Lefèvre, L., Gallego, M. C., Vaquero, J. M., Clette, F., Bravo-Paredes, N., ... Bautista, M. L. (2018, December). A Sunspot Catalog for the Period 1952 - 1986 from Observations Made at the Madrid Astronomical Observatory. *Solar Physics*, *293*(12), 164. doi: 10.1007/s11207-018-1387-1

Aparicio, A. J. P., Vaquero, J. M., Carrasco, V. M. S., & Gallego, M. C. (2014, November). Sunspot Numbers and Areas from the Madrid Astronomical Observatory (1876 - 1986). *Solar Physics*, *289*(11), 4335-4349. doi: 10.1007/s11207-014-0567-x

Baranyi, T., Győri, L., & Ludmány, A. (2016, November). On-line Tools for Solar Data Compiled at the Debrecen Observatory and Their Extensions with the Greenwich Sunspot Data. *Solar Physics*, *291*(9-10), 3081-3102. doi: 10.1007/s11207-016-0930-1

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. (2012, March). Toward Reliable Benchmarking of Solar Flare Forecasting Methods. *Astrophysical Journal Letters*, *747*(2), L41. doi: 10.1088/2041-8205/747/2/L41

Bornmann, P. L., & Shaw, D. (1994, March). Flare Rates and the Mcintosh Active Region Classifications. *Solar Physics*, *150*(1-2), 127-146. doi: 10.1007/BF00712882

Carrasco, V. M. S., Álvarez, J. V., & Vaquero, J. M. (2015, October). Sunspots During the Maunder Minimum from Machina Coelestis by Hevelius. *Solar Physics*, *290*(10), 2719-2732. doi: 10.1007/s11207-015-0767-z

Carrasco, V. M. S., Vaquero, J. M., Gallego, M. C., Lourenço, A., Barata, T., & Fernandes, J. M. (2018, November). Sunspot Catalogue of the Observatory of the University of Coimbra (1929 - 1941). *Solar Physics*, *293*(11), 153. doi: 10.1007/s11207-018-1373-7

Carvalho, S., Gomes, S., Barata, T., Lourenço, A., & Peixinho, N. (2020, July). Comparison of automatic methods to detect sunspots in the Coimbra Observatory spectroheliograms. *Astronomy and Computing*, *32*, 100385. doi: 10.1016/j.ascom.2020.100385

Clette, F., Berghmans, D., Vanlommel, P., Van der Linden, R. A., Koeckelenbergh, A., & Wauters, L. (2007). From the wolf number to the international sunspot index: 25 years of sidc. *Advances in Space Research*, *40*(7), 919-928. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0273117707001974` doi: https://doi.org/10.1016/j.asr.2006.12.045

Colak, T., & Qahwaji, R. (2008, April). Automated McIntosh-Based Classification of Sunspot Groups Using MDI Images. *Solar Physics*, *248*(2), 277-296. doi: 10.1007/s11207-007-9094-3

Comaniciu, D., & Meer, P. (2002). Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(5), 603-619. doi: 10.1109/34.1000236

Cortie, A. L. (1901, May). On the Types of Sun-Spot Disturbances. *Astrophysical Journal*, *13*, 260. doi: 10.1086/140816

Curto, J. J., Blanca, M., & Martínez, E. (2008, August). Automatic Sunspots Detection on Full-Disk Solar Images using Mathematical Morphology. *Solar Physics*, *250*(2), 411-429. doi: 10.1007/s11207-008-9224-6

Curto, J. J., Solé, J. G., Genescà, M., Blanca, M. J., & Vaquero, J. M. (2016, November). Historical Heliophysical Series of the Ebro Observatory. *Solar Physics*, *291*(9-10), 2587-2607. doi: 10.1007/s11207-016-0896-z

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Multiple Classifier Systems* (pp. 1–15). Berlin, Heidelberg: Springer. doi: 10.1007/3-540-45014-9_1

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (p. 226–231). AAAI Press.

Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, *21*(1), 32-40. doi: 10.1109/TIT.1975.1055330

Gallagher, P. T., Moon, Y. J., & Wang, H. (2002, September). Active-Region Monitoring and Flare Forecasting I. Data Processing and First Results. *Solar Physics*, *209*(1), 171-183. doi: 10.1023/A:1020950221179

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA, USA: MIT Press. (`http://www.deeplearningbook.org`)

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017, 06–11 Aug). On calibration of modern neural networks. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1321–1330). PMLR.

Győri, L., Ludmány, A., & Baranyi, T. (2017, February). Comparative analysis of debrecen sunspot catalogues. *Monthly Notices of the Royal Astronomical Society*, *465*(2), 1259-1273. doi: 10.1093/mnras/stw2667

Hale, G. E., Ellerman, F., Nicholson, S. B., & Joy, A. H. (1919). The magnetic polarity of sun-spots. *Astrophysical Journal, vol. 49, p. 153*, *49*, 153.
Hathaway, D. H. (2015). The solar cycle. *Living reviews in solar physics*, *12*, 1–87.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 ieee conference on computer vision and pattern recognition (cvpr)* (p. 770-778). doi: 10.1109/CVPR.2016.90
Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J., & Maier-Hein, K. H. (2021, February). nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211. (Number: 2 Publisher: Nature Publishing Group) doi: 10.1038/s41592-020-01008-z
Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 ieee conference on computational intelligence in bioinformatics and computational biology (cibcb)* (pp. 1–7).
Kiepenheuer, K. O. (1953). Solar Activity. In G. P. Kuiper (Ed.), *The sun* (p. 322).
Knyazeva, I., Rybintsev, A., Ohinko, T., & Makarenko, N. (2020, January). Deep-Learning Approach for McIntosh-Based Classification Of Solar Active Regions Using HMI and MDI Images. In *Advances in neural computation, machine learning, and cognitive research iii* (pp. 239–245). doi: 10.1007/978-3-030-30425-6_28
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc.
Künzel, H. (1960, August). Die Flare-Häufigkeit in Fleckengruppen unterschiedlicher Klasse und magnetischer Struktur. *Astronomische Nachrichten*, *285*(5), 271. doi: 10.1002/asna.19592850516
Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.
LeCun, Y., Bengio, Y., & Hinton, G. (2015, 521). Deep learning. *Nature*, *521*, 436. doi: 10.1038/nature14539
Lefevre, L., & Clette, F. (2014, February). Survey and Merging of Sunspot Catalogs. *Solar Physics*, *289*(2), 545-561. doi: 10.1007/s11207-012-0184-5
Loshchilov, I., & Hutter, F. (2019). *Decoupled weight decay regularization.*
Léger, J., Leyssens, L., Kerckhofs, G., & De Vleeschouwer, C. (2022, September). Ensemble learning and test-time augmentation for the segmentation of mineralized cartilage versus bone in high-resolution microCT images. *Computers in Biology and Medicine*, *148*, 105932. doi: 10.1016/j.compbiomed.2022.105932
Mathieu, S., von Sachs, R., Ritter, C., Delouille, V., & Lefèvre, L. (2019, November). Uncertainty Quantification in Sunspot Counts. *Astrophysical Journal*, *886*(1), 7. doi: 10.3847/1538-4357/ab4990
McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. (2018, June). Flare forecasting using the evolution of McIntosh sunspot classifications. *Journal of Space Weather and Space Climate*, *8*, A34. doi: 10.1051/swsc/2018022
McIntosh, P. S. (1990, September). The Classification of Sunspot Groups. *Solar Physics*, *125*(2), 251-267. doi: 10.1007/BF00158405
Muraközy, J. (2022, January). Variations of the Internal Asymmetries of Sunspot Groups during Their Decay. *Astrophysical Journal*, *925*(1), 87. doi: 10.3847/1538-4357/ac3de6
Nguyen, T. T., Willis, C. P., Paddon, D. J., & Nguyen, H. S. (2004). On learning of sunspot classification. In M. A. Kłopotek, S. T. Wierzchoń, & K. Trojanowski (Eds.), *Intelligent information processing and web mining* (pp. 59–68). Berlin, Heidelberg: Springer Berlin Heidelberg.
Nguyen, T. T., Willis, C. P., Paddon, D. J., & Nguyen, H. S. (2006). A hybrid system for learning sunspot recognition and classification. In *2006 international*

conference on hybrid information technology (Vol. 2, p. 257-264). doi: 10.1109/ICHIT.2006.253620

Owens, B. (2013, March). Long-term research: Slow science. *Nature*, *495*(7441), 300-303. doi: 10.1038/495300a

Palladino, L., Ntagiou, E., Klug, J., Palacios, J., & Keil, R. (2022, March). Sunspot Groups Detection and Classification on SDO/HMI Images using Deep Learning Techniques. In *2022 IEEE Aerospace Conference (AERO)* (pp. 1–10). (ISSN: 1095-323X) doi: 10.1109/AERO53065.2022.9843222

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). *PyTorch: An imperative style, high-performance deep learning library.* doi: arXiv:1912.01703

Peake, E. J., Chevasson, R., Pszczolkowski, S., Auer, D. P., & Arthofer, C. (2020). Ensemble learning for robust knee cartilage segmentation: data from the osteoarthritis initiative. *BioRxiv*, 2020–09.

Pettauer, T., & Brandt, P. N. (1997, September). On Novel Methods to Determine Areas of Sunspots from Photoheliograms. *Solar Physics*, *175*(1), 197-203. doi: 10.1023/A:1004903201224

Phillips, K. J. (1991). Solar flares: A review. *Vistas in Astronomy*, *34*, 353-365. doi: https://doi.org/10.1016/0083-6656(91)90014-J

Priest, E. R., & Forbes, T. G. (2002, January). The magnetic nature of solar flares. *Astronomy and Astrophysics Reviews*, *10*(4), 313-377. doi: 10.1007/s001590100013

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Sammis, I., Tang, F., & Zirin, H. (2000, September). The Dependence of Large Flare Occurrence on the Magnetic Structure of Sunspots. *Astrophysical Journal*, *540*(1), 583-587. doi: 10.1086/309303

Sayez, N. (2023). *SunSCC* [software]. Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.8236706` doi: 10.5281/zenodo.8236706

Scherrer, P. H., Bogart, R. S., Bush, R. I., Hoeksema, J. T., Kosovichev, A. G., Schou, J., ... MDI Engineering Team (1995, December). The Solar Oscillations Investigation - Michelson Doppler Imager. *Solar Physics*, *162*(1-2), 129-188. doi: 10.1007/BF00733429

Scherrer, P. H., Schou, J., Bush, R., Kosovichev, A., Bogart, R., Hoeksema, J., ... others (2012). The helioseismic and magnetic imager (hmi) investigation for the solar dynamics observatory (sdo). *Solar Physics*, *275*(1), 207–227.

Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., & Zhang, H. (2018, June). A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr) workshops.*

SOLARNET [dataset]. (2021). Royal Observatory of Belgium. Retrieved from `https://solarnet.oma.be`

SOON [dataset]. (2008). `https://www.swpc.noaa.gov/products/solar-region-summary`. National Oceanic and Atmospheric Administration.

Steinegger, M., Bonet, J. A., Vázquez, M., & Jiménez, A. (1998, January). On the Intensity Thresholds of the Network and Plage Regions. *Solar Physics*, *177*(1-2), 279-286. doi: 10.1023/A:1004920125173

Sugirtha, T., & Sridevi, M. (2022). Semantic segmentation using modified u-net for autonomous driving. In *2022 ieee international iot, electronics and mechatronics conference (iemtronics)* (p. 1-7). doi: 10.1109/IEMTRONICS55184.2022.9795710

Turmon, M., Pap, J. M., & Mukhtar, S. (2002, March). Statistical Pattern Recognition for Labeling Solar Active Regions: Application to SOHO/MDI Imagery.

Astrophysical Journal, *568*(1), 396-407. doi: 10.1086/338681
Uset. (1956). Retrieved from https://www.sidc.be/uset/
USET Database [dataset]. (2022). Royal Observatory of Belgium. Retrieved from https://vo-tap.oma.be/
Watson, F., Fletcher, L., Dalla, S., & Marshall, S. (2009, November). Modelling the Longitudinal Asymmetry in Sunspot Emergence: The Role of the Wilson Depression. *Solar Physics*, *260*(1), 5-19. doi: 10.1007/s11207-009-9420-z
Zharkov, S., Zharkova, V., Ipson, S., & Benkhalil, A. (2005, December). Technique for Automated Recognition of Sunspots on Full-Disk Solar Images. *EURASIP Journal on Applied Signal Processing*, *2005*, 318462. doi: 10.1155/ASP.2005.2573
Zharkova, V. V., Aboudarham, J., Zharkov, S., Ipson, S. S., Benkhalil, A. K., & Fuller, N. (2005, May). Solar Feature Catalogues In EGSO. *Solar Physics*, *228*(1-2), 361-375. doi: 10.1007/s11207-005-5623-0