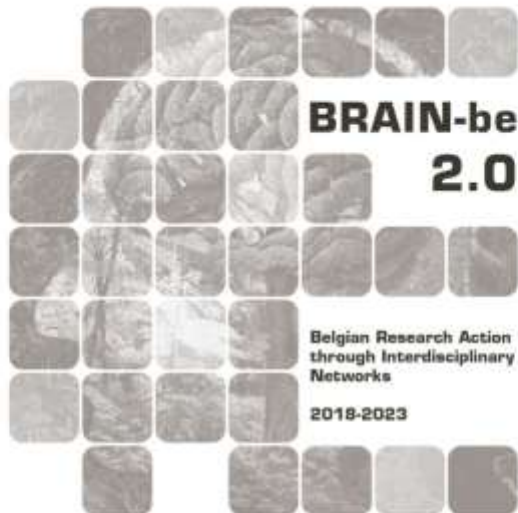


## **DELPHFI**

### **Deep learning prediction and hindsight of flare initiation**

Laurent Dolla (Royal Observatory of Belgium) – Jasmina Magdalenic Zhukov (ROB/KU Leuven University) – Panagiotis Gonidakis (KU Leuven) – Francesco Carella (KU Leuven) – Ekaterina Dineva (KU Leuven) – Philippe Vong (ROB/KU Leuven)

Pillar 1: Challenges and knowledge of the living and non-living world



NETWORK PROJECT

## DELPHI

Deep learning prediction and hindsight of flare initiation

Contract - B2/202/P1/DELPHI

## FINAL REPORT

**PROMOTORS:** Laurent Dolla (Royal Observatory of Belgium)  
Giovanni Lapenta (KU Leuven, deceased on 28/05/2024)  
Jasmina Magdalenić (KU Leuven, starting from 2024)

**AUTHORS:** Laurent Dolla (Royal Observatory of Belgium)  
Jasmina Magdalenić Zhukov (ROB/KU Leuven University)  
Panagiotis Gonidakis (KU Leuven)  
Francesco Carella (KU Leuven)  
Ekaterina Dineva (KU Leuven)  
Philippe Vong (ROB/KU Leuven)



Published in 2026 by the Belgian Science Policy Office  
WTCIII  
Simon Bolivarlaan 30 bus 7  
Boulevard Simon Bolivar 30 bte 7  
B-1000 Brussels  
Belgium  
Tel: +32 (0)2 238 34 11  
<http://www.belspo.be>  
<http://www.belspo.be/brain-be>

Contact person: Koen Lefever  
Tel: +32 (0)2 238 35 51

Neither the Belgian Science Policy Office nor any person acting on behalf of the Belgian Science Policy Office is responsible for the use which might be made of the following information. The authors are responsible for the content.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without indicating the reference:

Dolla, L; Magdalenic Zhukov, J; Gonidakis, P; Carella, F; Dineva, E; Vong, P; **DELPHI: Deep learning prediction and hindsight of flare initiation**. Final Report. Brussels: Belgian Science Policy Office 2026 – 61 p. (BRAIN-be 2.0 - (Belgian Research Action through Interdisciplinary Networks))

**TABLE OF CONTENTS**

<b>ABSTRACT</b>	<b>6</b>
CONTEXT .....	6
OBJECTIVES .....	6
CONCLUSIONS.....	6
KEYWORDS.....	6
<b>1. INTRODUCTION</b>	<b>7</b>
<b>2. STATE OF THE ART AND OBJECTIVES</b>	<b>8</b>
2.1 CONTEXT .....	8
2.2 STATE OF THE ART .....	8
2.3 OBJECTIVES OF THE DELPHFI PROJECT .....	9
2.4 ORIGINALITY WITH RESPECT TO THE RESEARCH FIELD.....	9
<b>3. METHODOLOGY</b>	<b>11</b>
3.1 DATA DESCRIPTION.....	12
3.2 FLARE PREDICTIONS APPLYING CNNs ON LINE-OF-SIGHT MAGNETOGRAMS OF INDIVIDUAL ACTIVE REGIONS.....	13
3.2.1 Data preprocessing	13
3.2.2 Description of the used CNN architectures	16
3.2.3 Evaluation metrics	19
3.2.4 Impact of image scaling methods on the flare forecasting process and abilities of a CNN	20
3.3 ACTIVE REGION PARAMETRIZATION WITH VARIATIONAL AUTOENCODERS .....	22
3.4 SOLAR FLARE PREDICTION USING CNNs ON FULL-SUN EUV IMAGES.....	23
3.4.1 Solar Coronal Structure Segmentation	27
3.4.3. Solar coronal segmentation on level-0 SDO data	29
<b>4. SCIENTIFIC RESULTS AND RECOMMENDATIONS</b>	<b>31</b>
4.1 FLARE PREDICTIONS APPLYING CNNs ON LINE-OF-SIGHT MAGNETOGRAMS OF INDIVIDUAL ACTIVE REGIONS.....	31
4.1.1 Evaluation of the classification of individual magnetograms	31
4.1.2 Monitoring of the time variation of the classification	34
4.1.3 Comparison of the results between architectures and loss functions	39
4.1.4 Comparison of the results between classifications systems	39
4.1.5 Analysis of results label-wise	41
4.1.6 Effect of image scaling processes of the data on the prediction performances	41
4.1.7 Regions of attention of the Neural Network	44
4.1.8 Limits of the current study	47
4.2 ACTIVE REGION PARAMETRIZATION AND CLASSIFICATION WITH VARIATIONAL AUTOENCODERS .....	48
4.2.1 Active region parametrization	48
4.2.2 Active region classification	49
4.3 FLARE PREDICTION USING CNNs ON FULL-SUN EUV IMAGES .....	51
4.4 RELATED RESULTS .....	51
4.5 RECOMMENDATIONS.....	51

4.5.1 Flare predictions applying CNNs on line-of-sight magnetograms of individual active regions	51
4.5.2 Solar flare prediction using CNNs on full-Sun EUV images and coronal structure segmentation	52
4.5.3 Active region parametrization and classification with Variational Autoencoders	52
<b>5. DISSEMINATION AND VALORISATION</b>	<b>53</b>
<b>6. PUBLICATIONS</b>	<b>56</b>
<b>7. ACKNOWLEDGEMENTS</b>	<b>57</b>
<b>REFERENCES</b>	<b>58</b>

## **ABSTRACT**

### **Context**

Forecasting solar eruptions is crucial for our modern society based on vulnerable technology.

### **Objectives**

Our objectives are to improve our understanding of the mechanisms that lead to solar flares and improve the flare forecast, thanks to modern Machine Learning techniques.

### **Conclusions**

We have designed Convolutional Neural Network models that are able to predict the occurrence of flares within the next 24 hours with a very good efficiency, from photospheric magnetograms. We show that it is essential to keep the native resolution and aspect ratio of the solar images, contrary to the habit in the field (that is driven by computing time and common neural network architecture requirements). This is achieved by using a special architecture. When input with line-of-sight magnetograms, the attention of the neural network is focused on the polarity inversion lines.

We have also produced a latent space representation of the magnetograms using  $\beta$ -Variational Autoencoders. We performed unsupervised clustering and structure analysis of the  $\beta$ -VAE latent features using methods such as t-SNE and HDBSCAN. We used a Long Short-Term Memory (LSTM) time-series model with two LSTM layers to predict the time evolution of the VAE latent dimensions, which encode key aspects of active region morphology. We then combined geometric/morphological features learned by the  $\beta$ -VAE with conventional SHARP parameters to predict flare occurrence. We also used k-nearest neighbours for flare now-casting.

Finally, we used Extreme Ultraviolet Images of the solar corona to predict the integrated soft X-ray flux in the next 30 minutes.

### **Keywords**

Solar Eruptions; solar Flares; soft X rays; magnetograms; Extreme Ultraviolet; Machine Learning; Deep learning; convolutional neural networks; Variational Autoencoders

## 1. INTRODUCTION

The Sun has a direct impact on our life and technology. The largest sources of space weather disturbances are the eruptions and flares produced by solar active regions (see examples of active regions in Figure 1). Their consequences on Earth can occur within a few hours after the initial flare, which is usually classified by its X-ray flux that can span 5 orders of magnitude. The impacted human activities are in demand of early warnings about the solar activity produced by space weather centres. However, the precise mechanisms that lead to the buildup of free energy in the solar corona and that trigger its release by magnetic reconnection are still unknown and a long-standing problem of solar physics.

Understanding and predicting flare behaviour is therefore a critical societal need. Our goal is to strengthen our understanding of the dynamics of solar flares using modern Machine Learning techniques. This will be done in a research-oriented approach, using “interpretable” machine learning techniques to understand the physics associated with solar flares, and high-definition satellite observations gathered during the past years by the Solar Dynamics Observatory mission.

This project is built around specific and innovative expertise of both the Royal Observatory of Belgium and the KU Leuven university: ROB hosts the Regional Warning Centre of the International Space Environment Service, in charge of producing daily space weather bulletins. The long-term goal of the project is to valorise our space weather forecasting services by improving the accuracy of the solar flare forecast. To achieve this goal, we work using an interdisciplinary approach, combining the expertise of ROB in solar data analysis and space weather, and KU Leuven experience in machine learning applications to heliophysics.

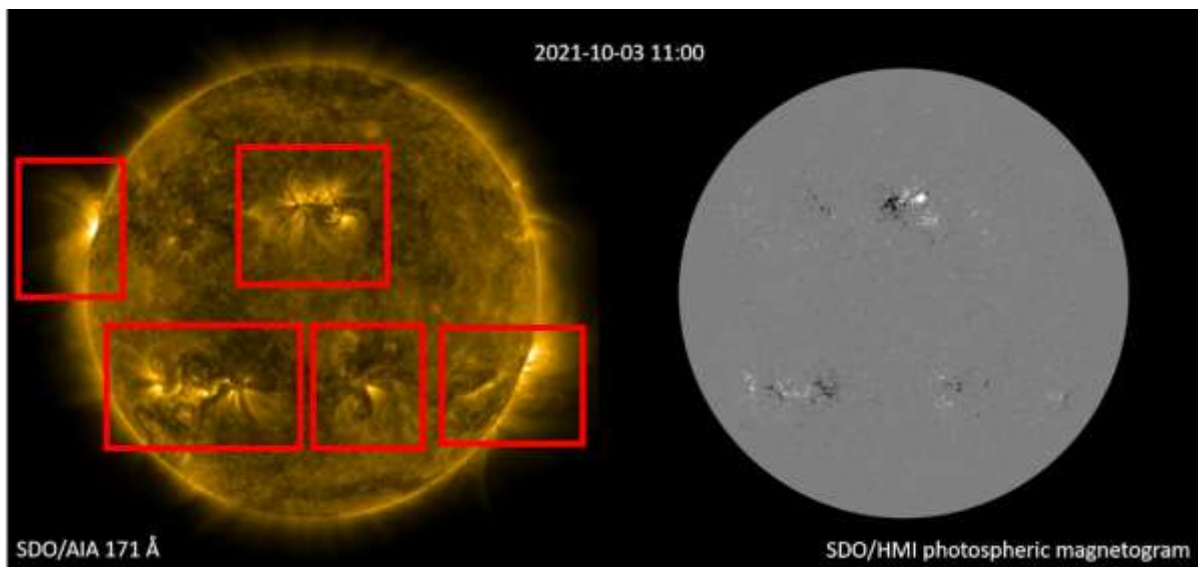


Figure 1: Example of solar active regions seen in different modalities on 2021/10/03. Left: EUV 171 Å (SDO/AIA), emitted by the solar coronal plasma at a temperature around 1 MK. The bright coronal loops trace some of the magnetic field lines. Right: photospheric magnetic field (SDO/HMI). The red rectangles delimit the active regions present on the Sun.

## 2. STATE OF THE ART AND OBJECTIVES

Active solar regions, associated to sunspots with high magnetic field activity are continuously tracked and catalogued. However, even today the accuracy of flare forecasting is still very low. In this project we propose to improve the flare forecasting accuracy by taking advantage of three major recent advancements: a) the availability of very high definition and continuous observations of the Sun in multiple wavelengths using the Solar Dynamics Observatory (SDO) mission, b) the maturity of modern Machine Learning techniques, and c) the affordability of high-speed processors and large capacity drives.

### 2.1 Context

Solar flares originate in the “active regions” of the Sun. These are sudden energy releases that emit extensive radiation in a very broad range of wavelengths from X-ray to radio ranges (Benz, 2017). They are usually classified by the X-ray flux radiation measured by instruments placed in the vicinity of the Earth. Such fluxes can span 5 orders of magnitude. Despite its importance, the precise mechanisms that lead to the buildup of free energy in the solar corona that trigger its release, by magnetic reconnection, are still unknown and a long-standing problem of solar physics (Shibata & Magara, 2011). Solar flares are also associated with large ejections of material from the Sun into the interplanetary medium (“coronal mass ejections”, CMEs) and with large scale flux of “solar energetic particles” (SEPs). Flare radiation, CME and SEPs can produce disturbances on the Earth electromagnetic environment (“space weather”; Schwenn, 2006). Some of these disturbances occur within a few hours following the original flare in the Sun. These include disturbances in radio communications, radar and Global Navigation Satellite System (GNSS), increase in the radiation of humans in the International Space Station, increase in the radiation dose of passengers in high altitude airplanes, and electric interference on spacecraft electronic components. All human activities are in demand of the early solar activity warnings provided by space weather centres.

### 2.2 State of the art

For the most part of flare prediction history, space weather warnings were based on a human interpretation of the state of the Sun. The earlier methods relied mostly on the shape and complexity of sunspots (the white light feature corresponding to active regions) to derive empirical relations for flare production (McIntosh 1990). More recently, the line-of-sight (LOS) and vector magnetic field data became the main source of information. A number of methods were based on the calculation of different physical features from solar magnetograms (magnetic flux, gradient of magnetic field, magnetic helicity, etc; Zirin & Wang, 1993; Leka & Barnes, 2003; Schrijver, 2007; Welsch et al., 2009). At the same time, some publications emphasized the temporal connection and evolution of different features as indicators for flaring activity (Wheatland, 2004; Falconer et al., 2012; Korsos, 2014, 2015).

The combination of growing data sets and improvements in computational power led gradually to the implementation of Machine Learning (ML) techniques for flare prediction (Yuan et al., 2010; Ahmed et al., 2013; Bobra & Couvidat, 2015; Nishizuka et al., 2017; Florios, 2018; for a comprehensive review, see Camporeale, 2019). Almost all of these ML methods are based on physical features derived from the photospheric magnetic field data (LOS and Vector). But several studies have shown that there is valuable information for the prediction of flares in the higher layers of the Sun's atmosphere (chromosphere, transition region, corona), emitting mainly in Ultra-Violet, Extreme Ultraviolet (EUV) or X rays (Nishizuka et al., 2017 and references within; Korsos et al., 2018; for complete reviews, see Fletcher et al., 2011, and Benz et al., 2017). Only a few studies have used information from that

category of data (Nishizuka et al., 2017; Jonas et al., 2018) and they are still only used to a limited extent. Recently, new studies that use Deep Learning, a subfield of ML, were published (Xin Huang et al., 2018; Nishizuka et al., 2018; Liu et al., 2019; Li et al., 2020; Wang et al., 2020). These have attracted a growing attention from the solar physics community.

Almost all the previous studies that use ML to predict solar flares are driven by active region properties built out of observational quantities. This list of (average or integrated) quantities are known as the active region features. Another approach, popular in the computer vision community, is to allow the learning algorithm to extract the features that describe the active regions directly from solar images. Convolutional Neural Networks (CNN) are a very prominent example of this approach.

Space weather prediction centres around the world use different techniques to produce forecasts, which range from fully automatic to forecaster in the loop (FITL) or only human predictions (see synopsis on Figure 1 in Leka et al., 2019). The results of a recent workshop comparing flare prediction performances of operational forecasting methods showed that none of them outperformed the rest by a significant margin, and only 4 out of 18 the methods used ML. None of them use Deep Learning (Leka et al., 2019 - papers II and III). Among the models evaluated, the current model of the Royal Observatory of Belgium (ROB) does not use any automatic method and only relies on human forecasts. It has been shown (Devos et al., 2014) that there is room for improvement, especially in comparison with basic ML methods.

### **2.3 Objectives of the DELPHI project**

1. To improve our understanding of the mechanisms that lead to solar flares, thanks to the interpretation of the results obtained by modern Machine Learning techniques;
2. To demonstrate how modern Machine Learning techniques, based on the automatic extraction of active region features, can provide better flare forecasts than human operators or existing automatic methods;
3. To increase the expertise of the Royal Observatory of Belgium in the development of Machine Learning flare predictions, and to set the ground for a new operational tool and its possible extension to other kinds of eruptive events (e.g. CMEs).

### **2.4 Originality with respect to the research field**

The innovative character of this project when it started, with respect to the existing research in the field of solar flare prediction, was to use some approaches that are rarely undertaken:

1. The flare prediction is done using modern CNNs and ML techniques that automatically extract features from solar images, which is contrary to other approaches where the active region features are handpicked by humans and pre-calculated. Advanced ML/CNN methods have hardly been used for flare prediction (see Xin Huang et al., 2018; Xuebao Li et al., 2020, for some limited and recent exceptions).
2. The project makes use of all available EUV solar disc observations from the SDO/AIA image, which contain information about the different temperature regimes of the solar atmosphere where reconnection energy is dissipated. We will couple this data with the routinely used vector magnetic field data. Most existing forecasting tools only use the photospheric magnetic field data, only a few studies include in addition EUV images in a few different wavelengths into their models (Nishizuka et al., 2017, Jonas et al., 2018).

3. The so-called eXplainable AI (XAI) techniques are used to derive a physical interpretation about the flaring initiation mechanism. This has been tried only once before (Xin Huang et al 2018).
4. For the first time in the domain of heliophysics “generative” methods, based on CNNs, are used to understand active regions, their properties and the conditions that lead to flares.
5. The project investigates the different parameters that have a strong impact on flare predictions, including their evolution in time, and their efficacy on different forecasting horizons.

### 3. METHODOLOGY

Flares are a manifestation of solar activity, when the increasingly twisted magnetic fields of the Sun have accumulated enormous amounts of energy in active regions, to the point of breaking the magnetic balance. Breaking the magnetic field lines is a process called magnetic reconnection. The problem is that it is impossible to have a direct observation of the magnetic field lines on the Sun. This is why we rely on secondary methods of measurement of the magnetic topology and the accumulation of energy in active regions. One of these methods reconstructs the magnetic field topology of active regions using 3D numerical models of the potential magnetic field, bounded by the vector magnetograms of the solar surface (a 2D surface). However, a potential field reconstruction makes the assumption that the magnetic field has no free energy stored, i.e. it does not reconnect.

A second method to assess the amount of stored free magnetic energy in an active region (called non-potential energy) is to use the vector magnetograms of the surface of the Sun and extract integral values of the mean magnetic field, mean magnetic helicity, and to construct the mean fluxes of current and magnetic flux across the surface. These quantities are routinely extracted by the SDO data pipeline, and stored in the Spaceweather HMI Activer Region Patch (SHARP) catalogue. However, this approach does not take into account the complex 3D structure of the magnetic field above the surface of the Sun.

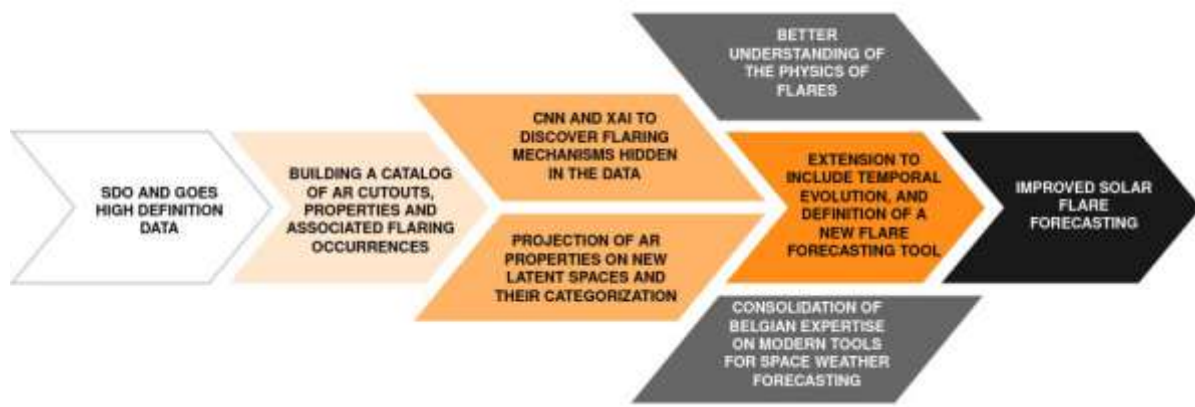


Figure 2: Overarching methodological approach.

Our methodological approach, as shown in Figure 2, is based on 4 main steps:

1. The collection and pre-processing of the solar data sets
2. The exploration and analysis of the predictive efficiency of CNNs on cutout magnetograms of the active regions
3. The exploration and analysis of the predictive information of the cutout magnetograms of active regions through the latent space representation produced by autoencoders
4. The exploration and analysis of temporal information in full-Sun EUV images to predict the total integrated soft X-ray flux.

The white arrow on the left of Fig. 1 shows the existing data available today: the SDO and the GOES missions have been gathering high-definition data for almost one solar cycle, including 4K (UHD) images of multiple wavelengths of the Sun, vector magnetograms, and X-ray fluxes, proxies of flare activity. We will take advantage of such a rich and large data set by implementing the procedures shown in orange in Figure 2.

### 3.1 Data description

The first and most critical step in our project is to gather the existing data and to build a catalogue of active regions (ARs) with enhanced information. Among the data products provided by the Joint Science Operations Center (JSOC) of the SDO mission, we can already find a comprehensive list of ARs detected by the HMI instrument. This list of HMI Active Region Patches (HARP) contains the coordinates of a bounding box (patch) that encompasses each AR. The HARP records all the patches as they cross the solar disk following the rotation of the Sun.



Figure 3: HARP entries on the full disk for the 13th of January 2013, taken from Bobra et al. (2014).

After detection of the HARP, the Space-Weather HMI Active Region Patches (SHARP) are extracted. The SHARP contains image cutouts of the vector ( $B_r$ ,  $B_\theta$ ,  $B_\phi$ ) and line-of-sight magnetic fields, velocity, and intensity (among other segments) for each HARP at each time step. It also includes keywords containing values of relevant space weather indices. In addition, the SHARP contains fifteen (15) space weather quantities derived from the parametrization of the photospheric vector magnetograms. SHARP records start in May 2010.

The SHARPs are produced every 12 minutes. Multiple AR can be present in the solar disk at any given time. The SHARP is then indexed by a timestamp and a unique record number.

The SHARP is a product from the Helioseismic and Magnetic Imager (HMI) instrument and does not contain cutouts of images coming from the Atmospheric Imaging Assembly (AIA) instrument. The AIA instrument produces continuous full-disk observations of the solar chromosphere and corona in ten (10) different wavelengths. These 4k (UHD) images are produced every 12 seconds.

The left panel of Figure 4 shows that active regions can present evolving features (moving from the top to the bottom row) in different wavelengths (columns) when a flare emerges.

We will also complement the active region entries with information about their flaring occurrence. This will be extracted from observations of soft X-ray fluxes by the Geostationary Operational Environmental Satellite (GOES) system.

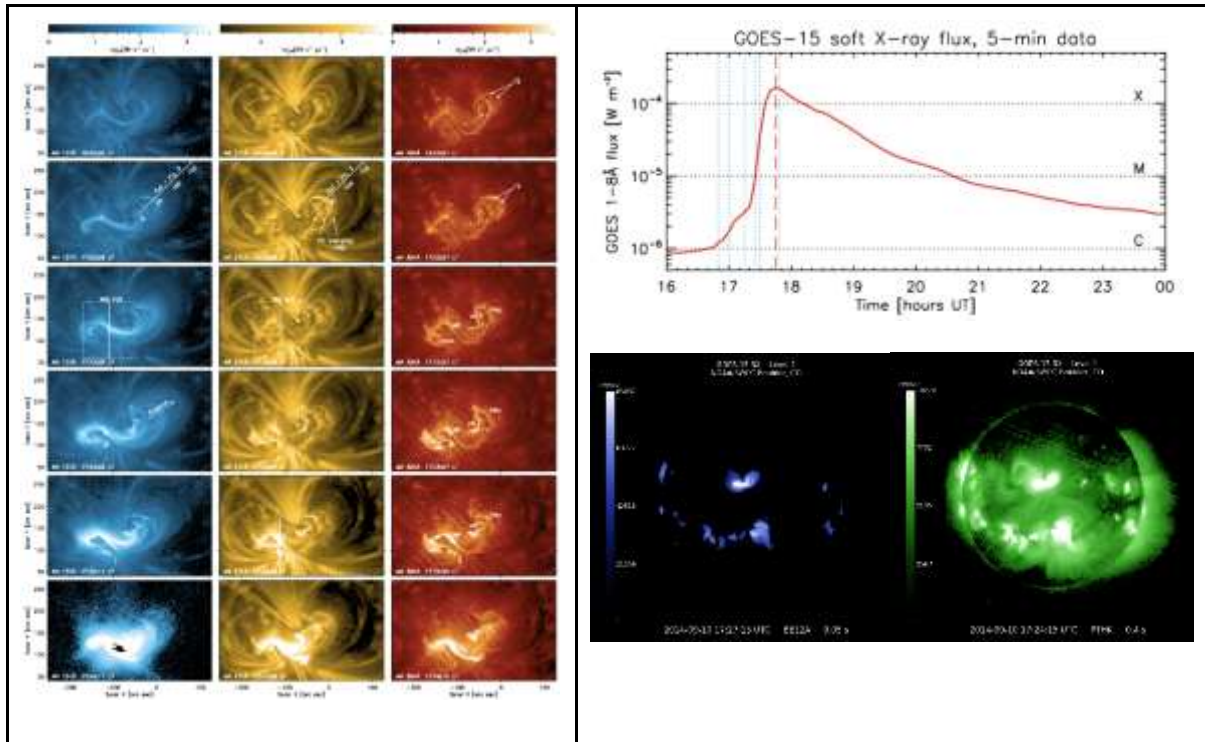


Figure 4: Evolution of an X1.6 flare, observed by the SDO/AIA instrument in three different wavelengths (left), and x-ray fluxes with full-disk x-ray imaging on the same date by the GOES mission (right), Dudík, J., et al. (2016).

### 3.2 Flare predictions applying CNNs on line-of-sight magnetograms of individual active regions

In this workpackage, we analyse the predictive potential of cut-out line-of-sight magnetograms of individual active regions (SHARP dataset) corresponding to the time interval from May 2010 to August 2021, over a forecasting horizon of 24 hours. It covers the end of the 24th solar cycle and the majority of the 25th. We couple the SHARP data with the GOES X-ray flare catalogue to label the images by using their common identification number of active regions provided by the National Oceanic and Atmospheric Administration (NOAA). The GOES catalogue is provided by the National Centers for Environmental Information (NCEI) and lists every solar flare observed since 1974 with their soft X-ray class, start time, end time, NOAA number, and many more parameters. We fetch it using the Heliophysics Events Knowledgebase (HEK) system through the Sunpy library (Sunpy community et al., 2020).

We compare the predictive efficiency of two different CNN architectures: one with a fixed input size (that we call Trad-CNN), and one that allows variable size input thanks to a special Spatial Pyramid Pooling layer (SPP-CNN).

#### 3.2.1 Data preprocessing

The data set is filtered. After removing SHARP files without an NOAA number (i.e., magnetic patches without a sunspot, and unlikely to produce any flare), 1579166 images remain in the dataset. If we used this whole dataset, the training of our models would last for months, especially with our limited resources. With this in mind, we decided to keep one image per hour. The SHARP dataset initially captures one image every 12 minutes. This selection reduces the dataset to 315374 images, which should be enough to produce a decent model without excessively long training.

Our dataset at this point presents images captured by HMI with no conditional selection. Unfortunately, a significant part of the dataset is unusable or may negatively impact the training of the model due to the size and position of the image. If we kept every image from the whole disk, there would be a consequential difference between SHARP files due to the projection linked to the curvature of the Sun. As such, we removed every image of active regions without the whole image within 45 deg from the central meridian. In addition to impacting the prediction performance of our models, the size of the image can also limit their architecture. Through the convolution and pooling processes of a CNN, the image can become too small to fit in the architecture of our models or lose all its information before the flattening layer. Acknowledging this problem, we decided to keep SHARP files with a width and height greater than 150 pixels. This leaves us with 141378 images representing 1233 active regions.

The dataset at this stage contains magnetograms cutouts (25417 images, 241 active regions) encompassing several smaller active regions (8902 images, 85 active regions). The metadata of those large cut-outs registers a list named NOAA\_ARS, enumerating the NOAA number of every region in the image. In the process of splitting the dataset into training, validation, and test dataset, there could be a larger active region with many smaller ARs in one dataset and magnetograms of the same smaller regions in another one. If this is not handled properly, this could compromise the parsimony of the dataset. To deal with this issue, we decided to remove every SHARP file when they are sub-regions of a larger one. Through all this filtering, the dataset now contains 132476 images and 1148 active regions.

To have a better visualization and control of its distribution, we labelled each image with seven distinct image labels:

- X: Images of an active region producing an X class flare in less than 24 hours.
- M: Images of an active region producing an M class flare in less than 24 hours.
- C: Images of an active region producing a C class flare in less than 24 hours.
- Future-X (FX): Images of an active region producing an X class flare in more than 24 hours.
- Future-M (FM): Images of an active region producing an M class flare in more than 24 hours.
- Future-C (FC): Images of an active region producing a C class flare in more than 24 hours.
- Never-Flare (NF): Images of an active region that will not produce any flare in the future.

Images with several flares in the following 24 hours will be labelled using the strongest flare.

For this study, we produced two different classifications. The first one is called CMX classification and includes C, M, and X images in the positive class, while the rest is in the negative class, and the second classification is called MX classification and only includes M and X images in the positive class. By training models on these two classifications, we could evaluate the difference in prediction performance for X and M classes between our classifications.

Furthermore, with this labelling system, we can produce the training, validation, and test datasets approximately similar to the real-life distribution inside each class (the training datasets are used during the training of our models, while the validation datasets are used to ensure their training is progressing correctly. The test dataset is used at the end to evaluate the prediction ability of our models). We first select the NOAA identification number of every active region with X-labelled images among the filtered dataset. We then shuffle this list and split them in an 8:1:1 distribution

corresponding to the training, validation and test sub-dataset. This distribution is then used to store the images with the corresponding NOAA number to their respective sub-dataset. Following this, we remove these images from the filtered dataset and proceed with active regions represented by the remaining M-labelled files. We continue this process with the remaining labels. The process order is the following: X, M, C, NF, FX, FM and then FC. This operation ensures that there is never the same active region between the training, validation, and test dataset, keeping the parsimony of the dataset and guaranteeing a correct distribution of flares per class.

By using this labelling system, we create a significant class imbalance due to the amount of X and M flares compared to C and Never-Flare images. Class imbalance is a major problem in DL and causes the models to prioritize the classification of the larger class and ignore the smaller one. Solutions to this problem are usually separated into two categories, called data augmentation and downsampling. Data augmentation is a method in computer vision to produce artificial training data by rotating, stretching, flipping, or using other means. This is usually a great way to tackle class imbalance while teaching models to recognize images in different forms. However, we cannot use this method to predict solar flares. Our images of active regions need to stay physically realistic based on our observation tools. Sunspots follow a logic such that by rotating an image for example, features such as the placement of negative and positive spots become unrealistic and if a model learns to recognize them, it may become confused during the testing phase and produce a lower prediction performance. However, downsampling is a possible solution to our problem. By randomly removing data from the majority class, we can balance the dataset at the expense of breaking the realistic flare distribution between classes. For the MX classification training dataset, we use downsampling with a 40:60 ratio, reducing the majority class data count to 1.5 times the quantity of data in the minority class. Although there is still a slight class imbalance, it is better than the previous 2:98 ratio. We decided not to lower the downsampling ratio anymore to keep a well-rounded distribution and enough data for the models to learn properly. If we chose to downsample to a 50:50 ratio, approximately 80 C-labelled images would have remained compared to 300 X-labelled images, which is not enough for a correct training. In the case of the CMX classification, we can use a 50:50 ratio due to the presence of the C-labelled images in the minority class, as it greatly increases the size of the dataset. If we used a greater ratio for the majority class, the training time would have increased with minimal improvement in the prediction performance of our models. The downsampling ratio for models trained using a Score-Oriented Loss function is adjusted to a 20:80 distribution as the SOL function helps to reduce the impact of the class imbalance issue and allows us to reduce the use of the downsampling method to use more data for our models training. Together with the downsampling method, we use a class-weight method, which increases the importance of correctly classifying the minority class for the model during its training, based on the data distribution between classes. This contributes to compensate for the class imbalance. The formula used to compute the class-weight for the  $c$  class is  $Weight(c) = n_{samples} / (n_{classes} \times n_{samples\_of\_class(c)})$ .

To fit the Trad-CNN input, we resize the magnetograms to 512×512 pixels using bi-linear interpolation through the TensorFlow resize function. 128×128, 100×100, and 198×198 are popular resize size choices. However, we chose this resize size to reduce the information loss in strong solar flares. Active regions with intense flares in the next 24 hours tend to be of a bigger size than low-activity active region. If we resized them to a tenth or a fifth of their size, the magnetograms would lose a lot of their details and there are strong indications that flare productivity is related to small scales features (e.g.,

for features that can be observed in line-of-sight magnetograms: strong gradients around the polarity inversion line (Jing et al., 2006), or the so-called “magnetic channel” (Wang et al., 2008) and “magnetic tongue” (Poisson et al., 2016).

In the case of SPP-CNN models, there should be no need for resizing. However, to train our models, we use the mini-batch method. By feeding multiple images at a time to the model during its training and calculating the mean gradient to optimize its weight, the model can converge faster to its optimum state while offering more stable progress during its training. However, this method requires a common size for inputs among a mini-batch. To respect this condition while taking advantage of the SPP layer, we group images in 10 groups based on the sum of their width and height. Each group corresponds to an interval of 400 pixels. The first group contains images with the sum of their width and height higher than 300 pixels and less than  $300+400=700$  pixels. The interval of the second group is between 700 and 1100, and we continue for each group. During the training, mini-batches are constructed by selecting images of the same group and then resized to the average size of the mini-batch. This process allows us to minimize the amount of resizing and therefore the loss of details. There was the possibility to feed our model images one by one but after many tests, we concluded that models delivered a better prediction performance with mini-batch while some models without this method were not able to learn at all. During the trainings of our models, we used mini-batch size of 16 images.

Magnetograms can present NaN values. This is either due to the image showing beyond the border of the Sun or artifacts appearing during its production. To avoid the spread of NaN values during the backpropagation, we replace them with zeros (0 is actually the average value of magnetograms, so this does not introduce much disturbance). Moreover, we normalize the input with a z-score normalization (Lecun et al., 2012) from the SciPy library before feeding them to the CNN.

### **3.2.2 Description of the used CNN architectures**

In the current state of computer vision, CNNs are a popular approach to recognize and classify images. Through multiple layers of data processing, they can automatically extract complex features from their inputs and then output a prediction by feeding them to fully connected layers. The main layers of CNNs (Yamashita et al., 2018) are the following:

1. Convolution layers are composed of input data, a convolutional window, and an output called a feature map. Their main goal is to extract features from their input by applying a convolution. This process can be described as adding each element of the image to its local neighbours weighted by the convolution window. A convolution layer is illustrated by the number of filters, a filter size, a stride number, and a padding mode.
2. Activation layers are usually used to apply nonlinear functions, such as rectified linear unit (ReLU; Agarap, 2019), softmax and sigmoid, to their inputs.
3. Pooling layers are built the same way as convolution layers except that their goal is to extract the maximum value, or average depending on the use, in the pooling window. They are designed to reduce the input dimensionality while keeping the most important features.
4. A single flattening layer is used to transform its input, which is usually feature maps, from an n-dimensional matrix to a one-dimensional array. This allows the data to fit in the input of the first fully connected layer.

5. Fully connected layers are the last step of a convolutional neural network. They are composed of neurons and each neuron of the current layer is connected to every neuron of the previous layer with weights between them. The last fully connected layer is composed of  $n$  neurons, where  $n$  is the number of classes in case of a number of class greater than two. For a binary classification, using a single output neuron associated with a threshold value between the two classes is enough to classify the data.

A traditional CNN is composed of a sequence of convolution layers, activation layers, and pooling layers. The end of the network is constituted of fully connected layers. The transition from one layer to another is based on parameters such as weights and bias and can amount up to billions depending on the size of the network. These parameters are fine-tuned by learning and adapting to the data by using gradient-based optimization and other methods. This is an important step to teach a neural network to classify an image, and it is referred to as training. The training of our models was composed of three main processes:

1. The first step is called forward propagation. We first feed pre-labelled data from a training dataset to the model one by one. It will try to predict their class and based on the output of the network, an error value will be calculated with an error function. Error functions, or loss functions, are chosen based on the type of data, the classification method, and many other parameters. Popular error functions include binary and categorical cross-entropy and mean squared error (MSE).
2. The second step is called backpropagation. We fine-tuned the weights of our various layers by using the error value. We use the chain gradient rule to calculate the gradient of each weight in the network from the last layer to the first, and then update them.
3. After the whole training dataset was fed to the model through forward propagation and backpropagation, we used the validation dataset to track the prediction performance of our model and the training efficiency. This also allows the automatic adjustment of some hyperparameters of the model for an optimal training. If there is no error and the model keeps learning, we continue the training of our model by shuffling the training dataset and repeat the previous process several times.

Each cycle of these three processes is called an epoch. Whenever a fixed number of epochs is achieved or when the metrics used to evaluate the efficiency of the prediction of our models stagnate, the training is stopped. This set of epochs forms one fold of a  $k$ -fold cross-validation. For each fold, we keep the best model produced among all epoch by selecting the model with the highest sum of validation precision and recall. The precision and recall are metrics used to evaluate the prediction performance of a model, and they will be explained at the beginning of section 4. This method allows us to select the best model to predict solar flares accurately and without over predicting them. Following this, we evaluate the selected models with their test datasets to ascertain their true prediction performance.

As mentioned, a CNN has several hyperparameters that need to be fixed beforehand or adjusted during the training such as the learning rate, the momentum, the loss function and the mini-batch size. These hyperparameters are crucial to train a model correctly as they can significantly impact its learning process. For instance, with a learning rate too low, the model may learn too slowly or never converge. However, the model may never reach its optimal state if the learning rate is too high.

Another variable aspect that can have a significant impact on the prediction ability is the architecture of the model. There is an infinite combination of layout and layers and the data can become over-transformed, unusable, or under-processed and too complex if the wrong architecture is used.

During this study, we have trained and analysed several model architectures before choosing which one to use. Each had varying layer composition and hyperparameters, such as implementing or removing batch normalization layer (Ioffe and Szegedy, 2015) and dropout (Srivastava et al. 2014). We fine-tuned the learning rate, the mini-batch size, convolutional layers feature window strides and sizes, number of layers and many more hyperparameters and elements of the architecture to obtain models with significant flare prediction performance. The differences between all architectures were sometimes minimal, but each modification brought a lot of variation in the prediction performances of our models. The model architectures and hyperparameters used in this paper were chosen based on the average prediction performance and stability across a 5-fold cross-validation.

In this study, we analyze and discuss the result of two model architectures that stood out while searching for the best architecture for our study. They are described in Figure 5 (see more details in our publication Vong et al., 2025). The first architecture we analysed is referred to as Trad-CNN, with fixed input size. The second architecture, SPP-CNN, is the same as Trad-CNN but swaps the position of the batch-normalization layers with the activation layers. In addition, it sets the stride of convolutional layers to one and implements an SPP layer (He et al., 2014) before the flattening layer instead of a max pooling layer. A spatial pyramid pooling layer is a layer that slices its input into a fixed number of tiles based on its parameter and selects the maximum value in each tile. This process produces a fixed input for the next layer and is especially interesting before the flattening layer, as fully connected layers are the only ones that need a fixed input. The SPP layer allows us to input images of any size in the models and to retain all their details otherwise lost due to the resizing process.

The SPP layer allows us to use inputs of different size, which is why there is no input size specified in the description of SPP-CNN. However, this advantage also brings a problem in the form of the minimum size of the input. As previously mentioned, we set the stride of our convolution layers to 1, which is to prevent images from disappearing. If we used the same parameters as the Trad-CNN architecture, we would be forced to exclude images smaller than 300 pixels in either width or height due to the data size reduction induced by the convolution and max pooling layers. The output of these layers was calculated with the following equation:  $Output_{edge\_size} = (W-K+2P)/S+1$ , where  $W$  is the width or height of the input,  $K$  the window size,  $P$  the padding and  $S$  the stride. With a convolution window stride of 1, we can restrict the exclusion of images to images with edge size lower than 150 pixels in width or height. This exclusion was done in datasets of both model architectures, during the data filtering process.

Another difference between the Trad-CNN and SPP-CNN architectures is the swap of order of the batch-normalization and activation layers. We found that the models based on the SPP-CNN architecture cannot be trained effectively if their batch-normalization layers are placed prior to the activation layers. The models ended up predicting every input as either zero or one. The exact reason is unknown and deserves further studies in future studies.

In this study, we also compare the results of our models trained using two distinct loss function: the Binary Cross-Entropy (BCE) and a Score-Oriented Loss (SOL; Marchetti et al., 2022) function optimizing the True Skill Statistics (TSS; Bloomfield, 2012) metric. The BCE is a widely used loss function while the

SOL, aiming to optimize a given score (or metric), is not available in the Keras libraries in opposition to the BCE. In this study, we implemented the SOL function manually while not knowing that it is available on GitHub (<https://github.com/cesc14/SOL>). In our case, the TSS is an interesting score to optimize due to its insensitivity to class imbalance. As the SOL function helps reduce the impact of the class imbalance problem, we trained our models, using this loss function, with a downsampling ratio lower than when using a BCE loss function. Reducing the downsampling ratio while using other class imbalance solutions increase the diversity of our dataset and can help The optimizer used during trainings is Adam, and its parameters are the default ones used in the Keras library. Every convolution in both architectures use the padding method “same.”

The models of this study were developed mostly using Python with the NumPy, Pandas and TensorFlow libraries and processes were run using an A100 graphical processing unit.

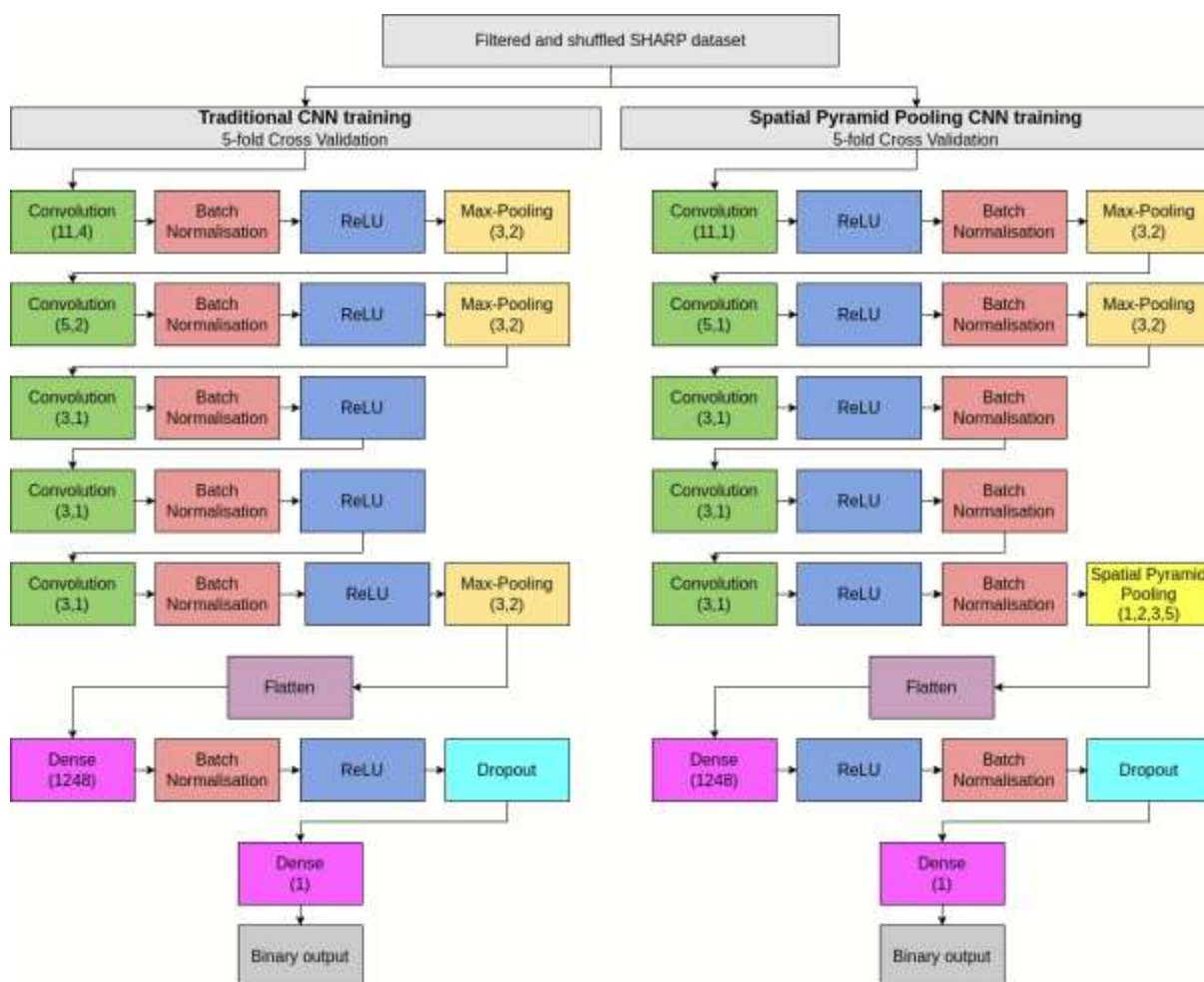


Figure 5: Diagram of Trad-CNN and SPP-CNN architectures used in this study.

### 3.2.3 Evaluation metrics

The models of this study output a binary classification. For the CMX classification, an image is classified as positive if it presents an active region that is flaring in less than 24 hours, or considered negative in any other case. In contrast, the MX classification only considers X and M labelled images as positive.

Based on the label of an image and its prediction by the model, we can deduce four possible outcomes. A prediction can be:

1. A true positive (TP), which means the prediction of the model and the label of the image are positive
2. A true negative (TN), which means the prediction of the model and the label of the image are negative
3. A false positive (FP), which means the prediction of the model is positive, but the label of the image is negative
4. A false negative (FN), which means the prediction of the model is negative, but the label of the image is positive

With these four outcomes, we can evaluate a model by making it predict the label of every image from a test dataset and build the confusion matrix which enumerates how much TP, TN, FP, and FN were output. From this matrix, it is then possible to calculate the characteristics of our models. We use:

1. The recall, which shows the proportion of images with a positive label correctly predicted:  $\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$ .
2. The precision, which shows the proportion of images predicted positive with a positive label:  $\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$ .
3. The accuracy, which shows the proportion of correct prediction:  $\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN})$ .

As much as these three characteristics present useful information, they are prone to class imbalance. Therefore, we focus our analysis on the Precision-Recall Area Under Curve (PR AUC) and the True Skill Statistics, which are less sensitive to this problem and more relevant to our case. The PR AUC is produced by computing the Precision-Recall Curve and evaluating the area under this curve while the TSS is calculated by the following formula:  $\text{TSS} = \text{TP}/(\text{TP}+\text{FN}) - \text{FP}/(\text{FP}+\text{TN})$ .

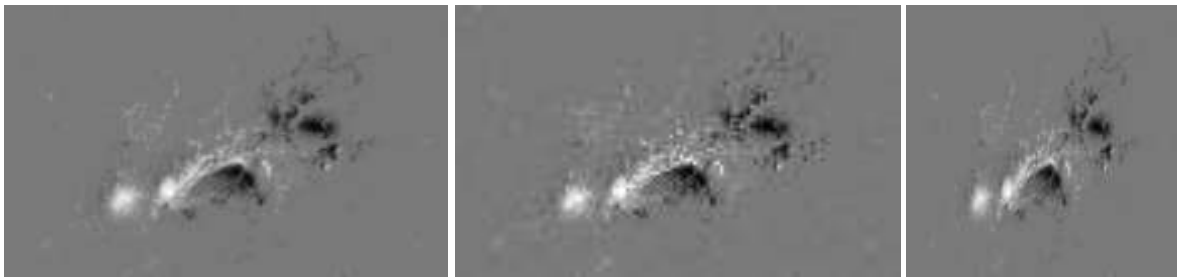
The trainings of our models were split into several epochs during which each model was trained on the whole training dataset and then evaluated on the validation dataset. Throughout these epochs, the validation accuracy, recall, and other metrics of our models varied based on how much the models were underfitting or overfitting. At the end of each training, the model with the best sum of validation-recall and validation-precision among the different epochs was kept and then evaluated using the corresponding test dataset.

### **3.2.4 Impact of image scaling methods on the flare forecasting process and abilities of a CNN**

This study aims to evaluate the impact of image scaling processes widely used in computer vision (in case of fixed size input images) on the predictive ability and attention of deep-learning models trained for flare forecasting. For this purpose, we use the SPP-CNN architecture (accepting input images of any size), and we evaluate the evolution of the predictive performances of the model using images scaled to various amplitudes. The image scaling processes we study are the resizing process, which modify the size of the image without impacting the aspect ratio of the image, and the stretching process, which changes the aspect ratio by stretching or compressing a single dimension of the image (see examples in Figure 6). The images used in this study are part of the Spaceweather HMI Active Region Patch (SHARP) line of sight magnetograms database.

In addition to the evaluation of the impact of image scaling processes on the predictive ability of the model, we also study its impact on the attention of the model and its method of prediction using the Gradient-weighted Class Activation Mapping method. This allows us to reassert a hypothesis expressed in previous studies about the importance of polarity inversion lines (PILs) for flare forecasting using deep-learning methods by quantifying the relation between the PILs and the regions of attention of the model during a prediction.

To realize this study, we used the test sub-datasets of the previous studies, as images of the test sub-dataset have not been used during the training of the respective models. Using this sub-dataset avoids the bias the models have with the training and validation sub-datasets, as well as reflects the use of novel data that may occur during an operational use of the model.



*Figure 6: Examples of resizing and stretching of an image. The leftmost image is the original image. The middle image has been resized to 20% of its original size, showing pixelisation. The rightmost image has been horizontally compressed to the size of its height (making it square).*

The amplitude of the image scaling processes is based on the resize and stretch ratios. For this study, we used a range of ratios from 0.2 to 2, with 0.05 steps. After scaling the dataset, we evaluated the prediction performance of our models on each of those scaled sub-datasets. We then used metrics output from the evaluation to compare the prediction performance based on the image scaling method and amplitude.

With the improvement of explainable artificial intelligence (XAI) methods, we are able to better understand the process of prediction of deep-learning models by looking at where the attention of the models is during the prediction. A prominent method of XAI is the Gradient-weighted Class Activation Mapping (Selvaraju et al., 2016). Using the output of the last convolution layer of a CNN and by back-propagating the gradients from the output of the model to the previously mentioned layer, we are able to obtain a heatmap highlighting the regions of the image which were important for the prediction. The Gradient-weighted class activation mapping (Grad-CAM) method allowed us to first compare the difference of attention between Trad-CNN and SPP-CNN models during a prediction. We believe this comparison to be significant in explaining the differences in metrics between the models, and it may also give us an insight into how the models realize their prediction. Using the Grad-CAM method, we then compared the evolution of the attention of the model through various amplitudes of image scaling. This allows us to visualize the impact of image scaling processes on the attention of the model. As a secondary goal, we also quantified the importance of PILs for flare prediction using LOS magnetograms, with the output of the Grad-CAM method.

### 3.3 Active region parametrization with Variational Autoencoders

This project takes advantage of variational autoencoders (VAEs; Kingma et al., 2019) – probabilistic encoder-decoder neural networks, to learn compact, information-rich representations of solar active region (AR) images. The input for this model are SHARP vector magnetic field (VMF) maps, and specifically the radial component of the field  $B_r$  (see Figure 7), often used to study the magnetic feature variability in context of solar eruption precursors. A VAE maps each input  $B_r$  map to a distribution over latent variables rather than a single deterministic code, enabling the model to capture uncertainty and the intrinsic variability of the region’s magnetic field morphology. Using a convolutional neural network (CNN) architecture for both the encoder and decoder leverages the spatial locality and multi-scale structure present in magnetograms, allowing the network to extract hierarchical features such as polarity inversion lines, flux concentrations, and evolving texture patterns.

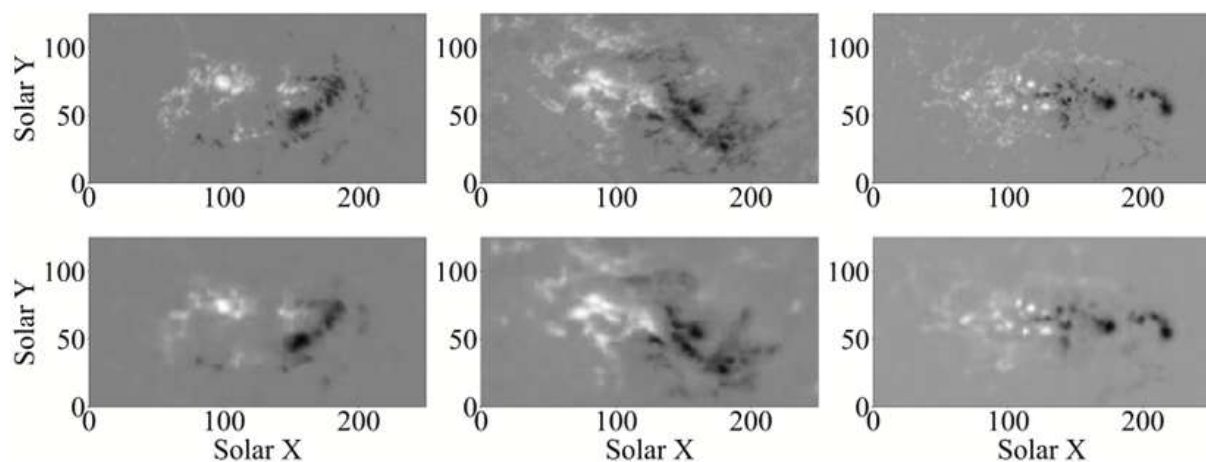


Figure 7: Testing the VAE: Comparison between original input (top) and generated output of the trained  $\beta$ -VAE model (bottom).

In a Variational Autoencoder (VAE), inputs are mapped to a continuous latent space via an approximate posterior that is explicitly regularized toward a simple prior by a Kullback–Leibler (KL) penalty, yielding a smooth, approximately factorized manifold on which linear interpolation is semantically meaningful and sampling from the prior produces coherent generative latent features. Applying the VAE model on the SHARP VMF maps, produces a latent representation which expresses the active region’s large-scale properties (e.g., overall morphology, size/scale, brightness/flux patterns, polarity layout). Furthermore, the VAE model is designed to produce a simple, smooth shared layout of the latent space, where interpolating between two points traces plausible transitions in AR appearance, and differences in position, which reflects genuine, dataset-level variations rather than idiosyncrasies of single images.

The disentangled VAE or  $\beta$ -VAE introduces a weighting coefficient  $\beta > 1$  for the KL term, which strengthens steering, pushing the representation to separate factors more cleanly so that individual latent features more often track distinct AR attributes (e.g., compactness, bipole separation, or texture complexity). This typically yields crisper clustering and more reliable cross-sample comparisons, useful for downstream analysis and time-series tracking. Note that increasing  $\beta$  can compromise fine details, producing smoother but less detailed reconstructions. The properties and

interpretability of the VAE- and  $\beta$ -VAE-encoded latent spaces for a selected sample of ARs were investigated.

To study how different VAE and  $\beta$ -VAE variants organize their embeddings, we project latent vectors into two dimensions using Principal Component Analysis (PCA; Jolliffe, 2002), t-distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten, 2014), and Uniform Manifold Approximation and Projection (UMAP; McInnes et al. 2018). PCA provides a linear baseline that highlights dominant variance directions and global trends, while t-SNE and UMAP emphasize local neighbourhood structure and cluster formation. Comparing these projections across models and  $\beta$  settings lets us assess changes in latent smoothness, separation, and interpretability (e.g., tighter grouping or clearer trajectories) relative to the standard VAE.

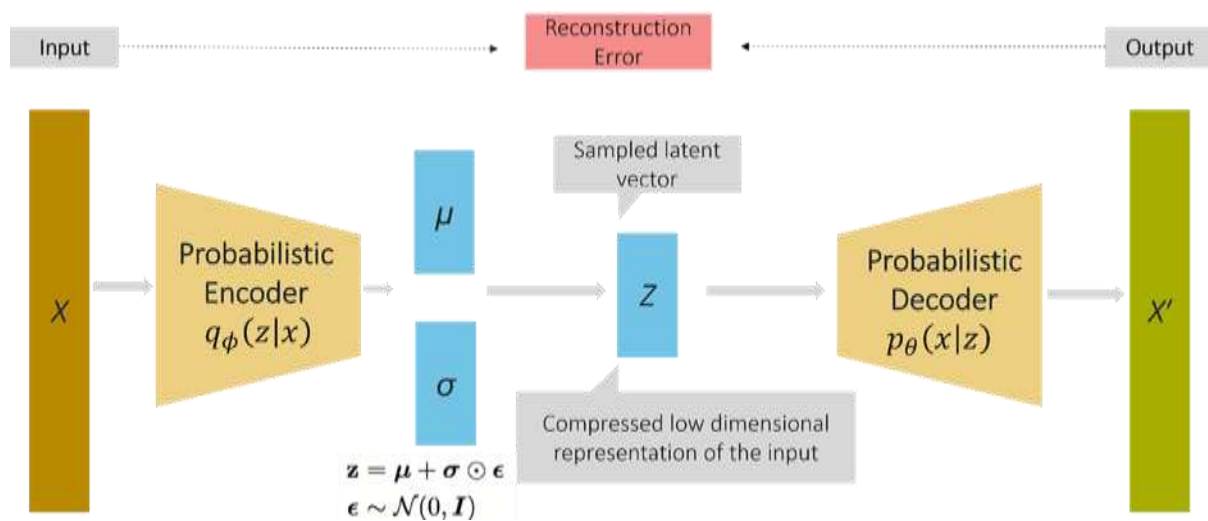


Figure 8: Schematic diagram of the VAE architecture.

### 3.4 Solar flare prediction using CNNs on full-Sun EUV images

The main goal of this task is to estimate the soft X-ray flux in the next hours from an EUV Full-Sun image. It also demonstrates the use of Deep Convolutional Neural Networks (DCNN) for the autonomous onboard analysis of solar images. The focus is on analysing high-resolution images of the Sun, particularly on the parametrisation of solar active regions.

We have identified that thresholding techniques applied to soft X-ray flux measurements can effectively detect solar flares. We aim to predict this soft X-ray flux using full disk images of the Sun obtained from the Solar Dynamics Observatory (SDO). We are collecting soft X-ray data from the NASA GOES Satellite to support this effort.

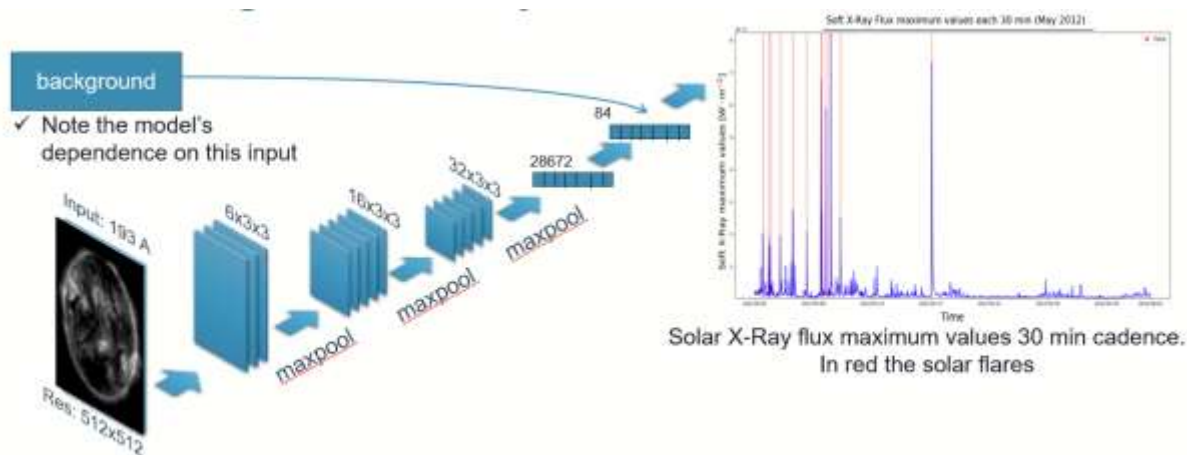


Figure 9: An example of the currently explored CNN architectures for soft X-ray flux prediction.

Our approach involves convolutional neural networks (CNNs), which are trained after proper data preprocessing. We have downloaded a 10-year dataset from SDO, specifically adapted for machine learning tasks, and have trained on this substantial data set. PyTorch and Optuna are being used to implement the CNNs and fine-tune their hyperparameters.

We utilize the *sdomlv2* dataset (Galvez et al. 2019), which provides downsampled and preprocessed SDO/HMI (magnetograms) and SDO/AIA (EUV) data, optimized for machine learning applications. The dataset's second version, which we employ, contains data spanning from 2010 to 2020. Initially, we worked with a smaller subset of the data to facilitate the development of our framework, focusing on samples from 2012. Our target variable is the soft X-ray flux, obtained from the NASA GOES satellite.

We conducted a thorough data examination to identify and remove duplicates or corrupted data. Additionally, data from various channels were temporally aligned, including the GOES soft X-ray flux, ensuring consistency across all inputs.

Afterwards, we carefully designed a data split that effectively quantifies model generalisation across the entire Solar Cycle 24. The proposed split is illustrated below.

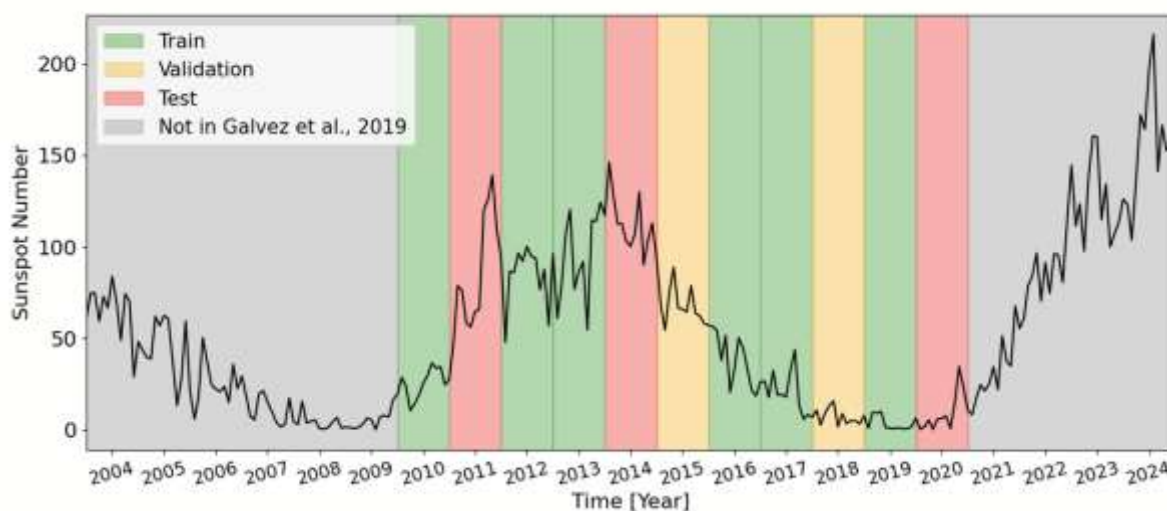


Figure 10: Proposed operational data split ensuring robust model generalisation across Solar Cycle 24.

Simultaneously, we have been incorporating a pretraining mechanism based on self-supervision, an approach previously introduced in biomedical engineering for segmentation and classification tasks (Zhou et al., 2021). More specifically, an encoder-decoder network is employed to reconstruct artificially deformed SDO images, following the self-supervision approach outlined in the Model Genesis algorithm.

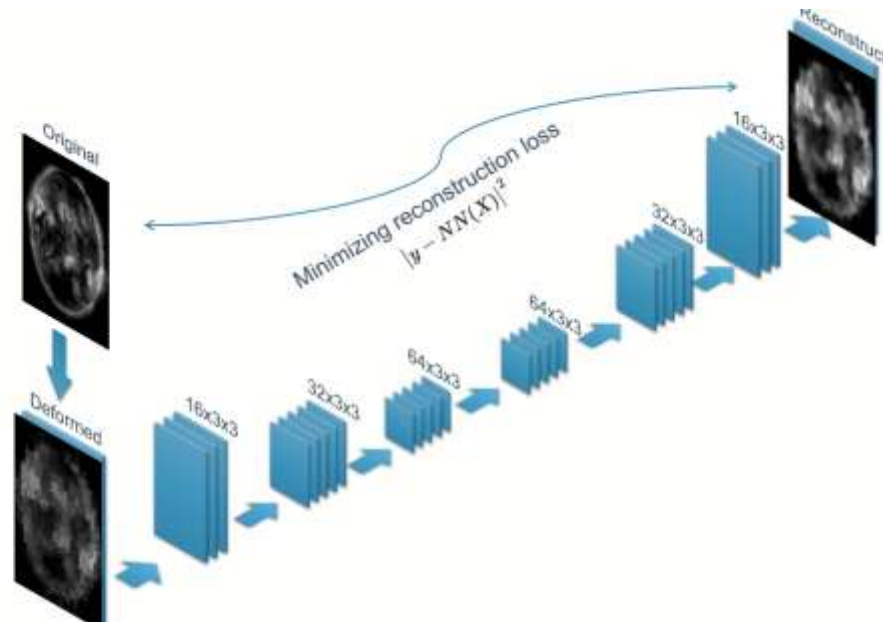


Figure 11: The self-supervision framework performing image reconstruction to derive an SDO/AIA pretrained model.

By learning to recover the original image from these deformations, the network is driven to extract meaningful and generalizable features from the data, improving its ability to understand complex structures in the solar images. This method is designed to provide an optimized initialization that is specifically tailored to our dataset, resulting in enhanced performance, improved data efficiency, and faster convergence during training. This self-supervised pretraining offers greater flexibility in network architecture compared to models pre-trained on ImageNet, enabling better adaptation to our unique input requirements. Previously reported results suggest that performance improvements can reach up to 10%, data efficiency can be increased by as much as 65%, and model convergence can be significantly accelerated, requiring over 50% fewer training epochs. However, we still need to assess whether these gains are applicable to our specific use case and dataset.

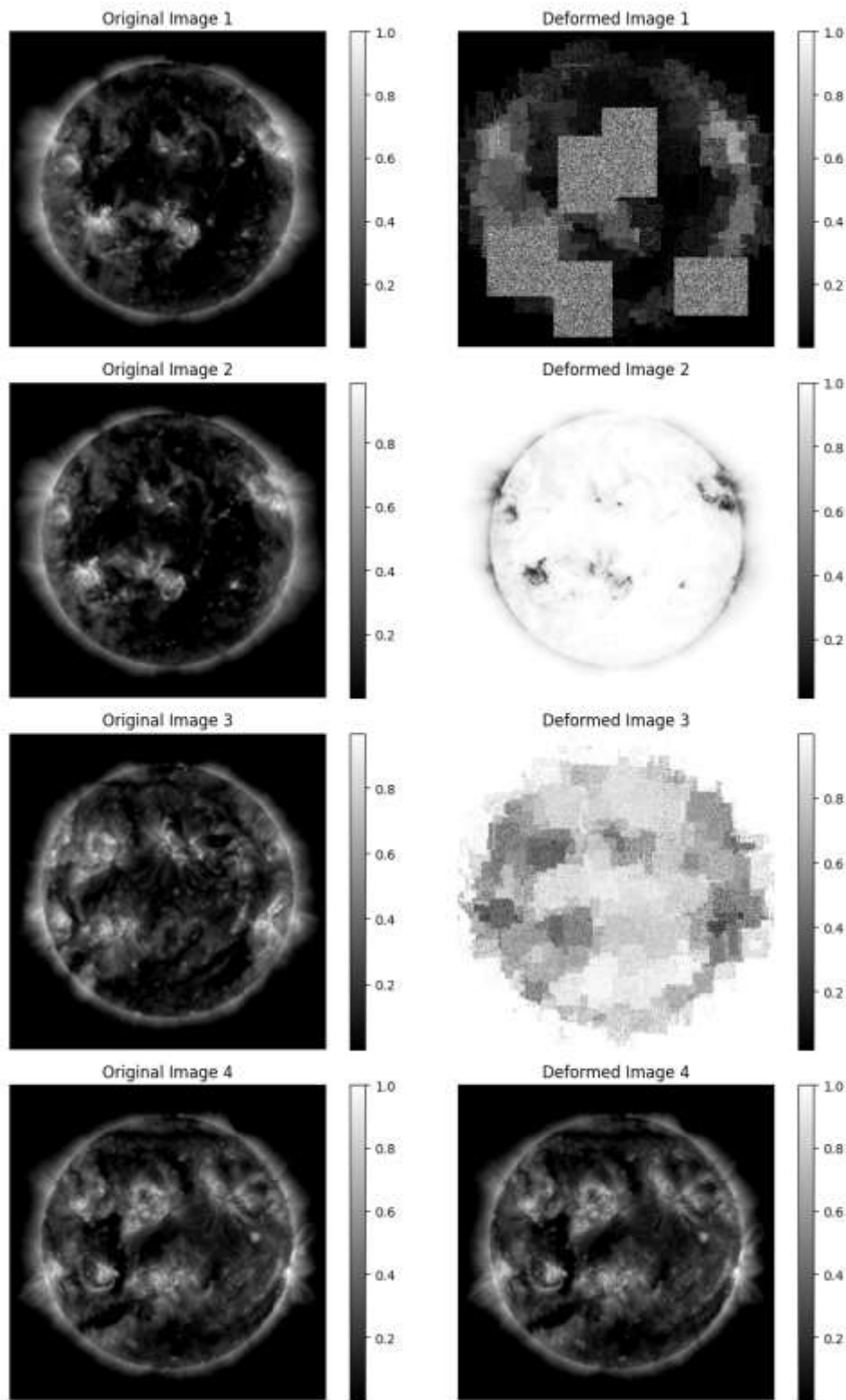


Figure 12: Training samples from the self-supervised pretraining: Right - original SDO images, Left - their artificially deformed counterparts.

### 3.4.1 Solar Coronal Structure Segmentation

Solar flare forecasting can be significantly improved when a coronal-structure segmentation framework is able to accurately extract the regions of interest from full-disk SDO images—primarily active regions and coronal holes. These segmented areas can then serve as inputs to a secondary processing pipeline designed to predict solar flares more reliably, similar to current forecasting methods that use HARP patches as input.

#### 3.4.1.1 Mackovjak et al. (2021) model based on a U-net-like architecture

Mackovjak et al. (2021) have recently introduced a new solar coronal structure segmentation mode for segmenting coronal holes and active regions using images from the Solar Dynamics Observatory's Atmospheric Imaging Assembly (SDO/AIA). This model adopts a U-net-like architecture, commonly used in biomedical image segmentation, ensuring precise delineation of solar features.

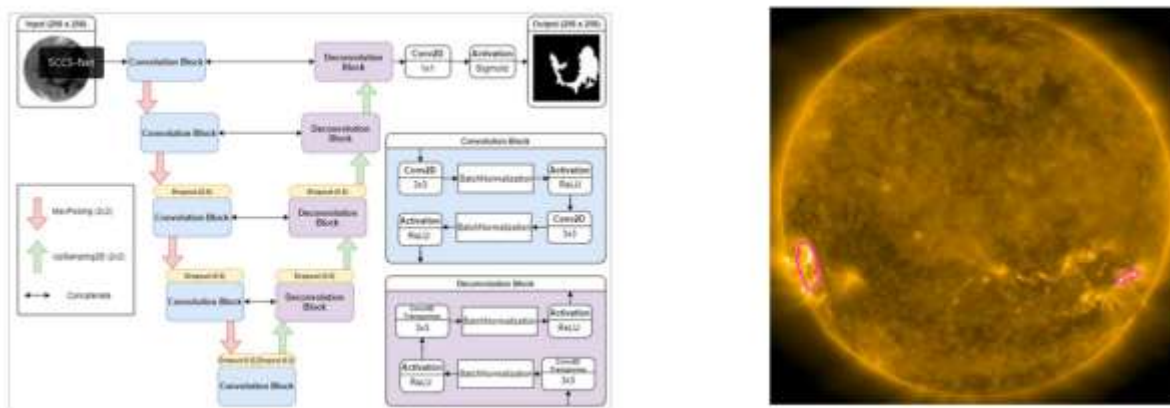


Figure 13: UNet-based architecture for solar coronal segmentation. Figure adapted from (Mackovjak et al. 2021).

The trained model is now publicly available and uses advanced image processing techniques to identify critical solar phenomena accurately. Its effectiveness is assessed using the Dice coefficient and the Intersection over Union (IoU) metrics, which confirm high accuracy and significant agreement with ground truth annotations.

The model uses manually annotated data covering active regions (AR) and coronal holes (CH) for training and validation. These annotations span various wavelengths, including 193 Angstroms (193 Å) and 171 Angstroms (171 Å), providing a range of conditions for robust model evaluation. Additionally, binary masks for both coronal holes and active regions are provided. These masks are essential for training the segmentation algorithms and allow users to benchmark their results against a standard dataset.

This innovative approach enhances the study and monitoring of solar activities and offers valuable insights into the dynamics of coronal holes and active regions through precise segmentation techniques.

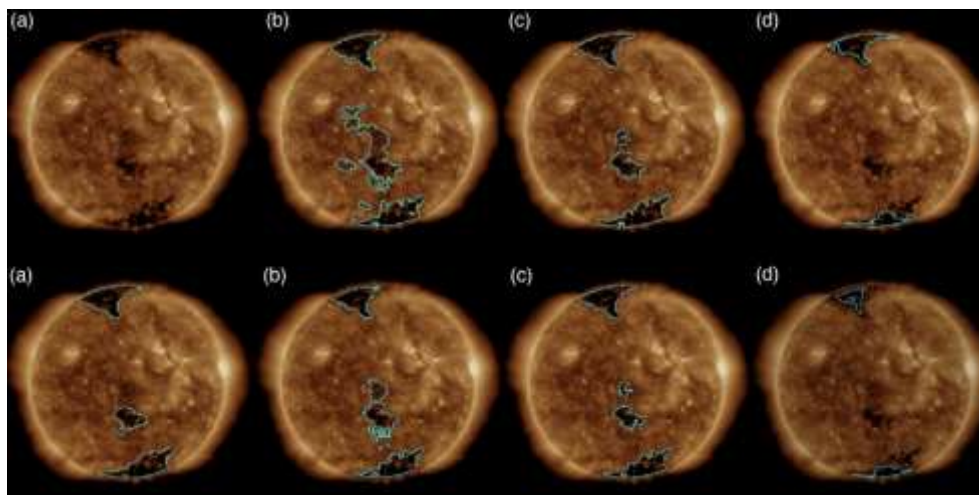


Figure 14: Examples of solar coronal structure segmentation. Figure adapted from (Mackovjak et al. 2021).

This approach, however, has several limitations. Firstly, due to its reliance on a U-Net-based segmentation network, it is relatively resource-intensive, making it less suitable for embedded systems and not applicable to be deployed on CPUs. Additionally, while the authors provide the code, the resulting trained model, and a limited number of evaluation samples, this scarcity of data complicates comprehensive quantitative evaluation.

### 3.4.2 Solar structure segmentation based on morphological filtering

As a result, we have been developing a similar but significantly lighter segmentation framework that utilizes morphological filters and thresholding (see Figure 15). Following segmentation, we can apply unsupervised clustering based on engineered features, enabling us to segment coronal structures and assign corresponding labels.

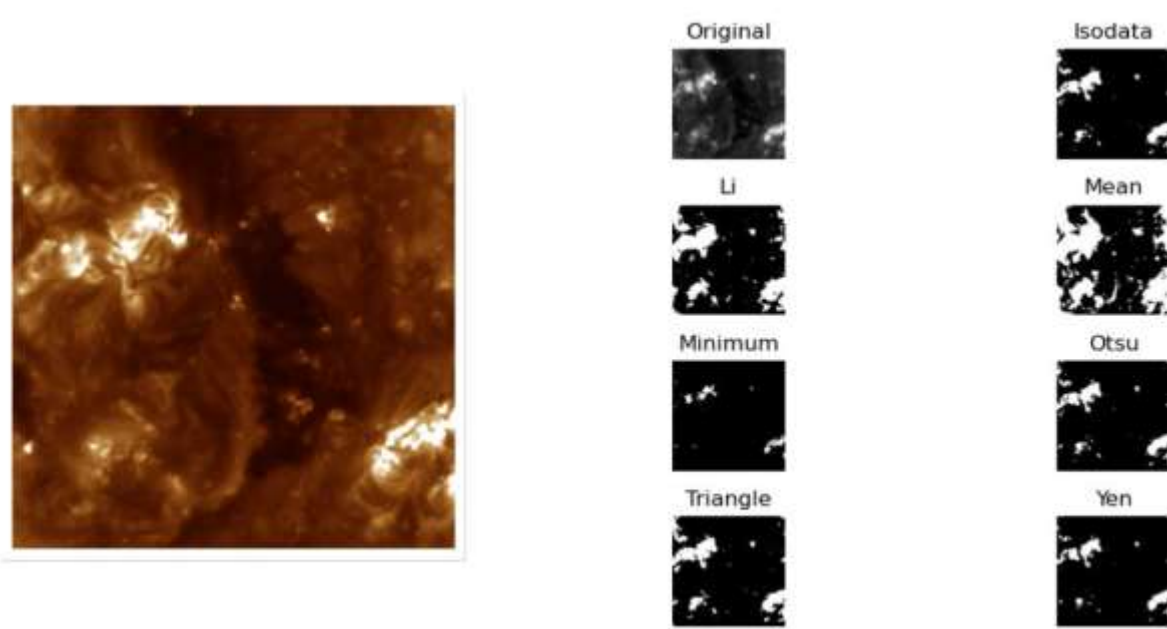


Figure 15: Examples of solar coronal structure segmentation using various thresholding methods. Ground truth masks were manually created to evaluate the performance of different thresholding algorithms, both through visual inspection and quantitative metrics, such as the Dice score.

### 3.4.3. Solar coronal segmentation on level-0 SDO data

Training and evaluation of the solar coronal structure frameworks are conducted on well-calibrated SDO data, often referred to as Level-2 data, as described in the previous sections. These data products have undergone extensive preprocessing, including calibration, alignment, despiking, and correction for instrumental effects such as instrument degradation and other systematic errors. As a result, Level-2 data provide clean, stable inputs ideal for supervised learning and model benchmarking. However, a major challenge arises when considering the onboard deployment of such models for future space missions: spacecraft typically have access only to Level-0 data, which are raw, uncalibrated telemetry streams directly from the instrument. These data include various instrumental artefacts, noise, and format differences that are absent from Level-2 products. Therefore, models trained on high-quality Level-2 data may fail or underperform when exposed to raw Level 0 inputs in operational settings. To bridge this gap, either onboard preprocessing pipelines must be developed to approximate Level-2 calibration, or models must be fine-tuned to become robust or be inherently generalisable to the noise and variability present in Level-0 data. Addressing this domain mismatch is essential for ensuring the reliability and autonomy of AI-powered systems deployed in space.

To this end, we acquired two subsets of level-0 SDO derived from 2024 and 2025. Then, we aimed to define a lightweight pre-processing pipeline to make them compatible with our machine learning models, which were trained on level-2 data. In our case, when applying a global normalisation to the  $[0,1]$  range, distinct quadrant discontinuities become apparent. These artefacts arise due to the independent acquisition and compression of each quadrant of the CCD sensor. Even after applying full-image normalisation combined with percentile-based clipping (1st and 99th percentiles), these discontinuities remain visible. To address this, we experimented with applying normalisation separately to each quadrant. While this approach occasionally reduces the artefacts, in many cases, residual discontinuities persist in one or more quadrants. As a final correction, we shift the median intensity of each quadrant to match the global median before performing normalisation. This method consistently suppresses the quadrant boundaries, resulting in visually continuous images across the whole frame. Figure 16 illustrates how the above preprocessing steps progressively address the quadrant discontinuities present in Level 0 SDO data.

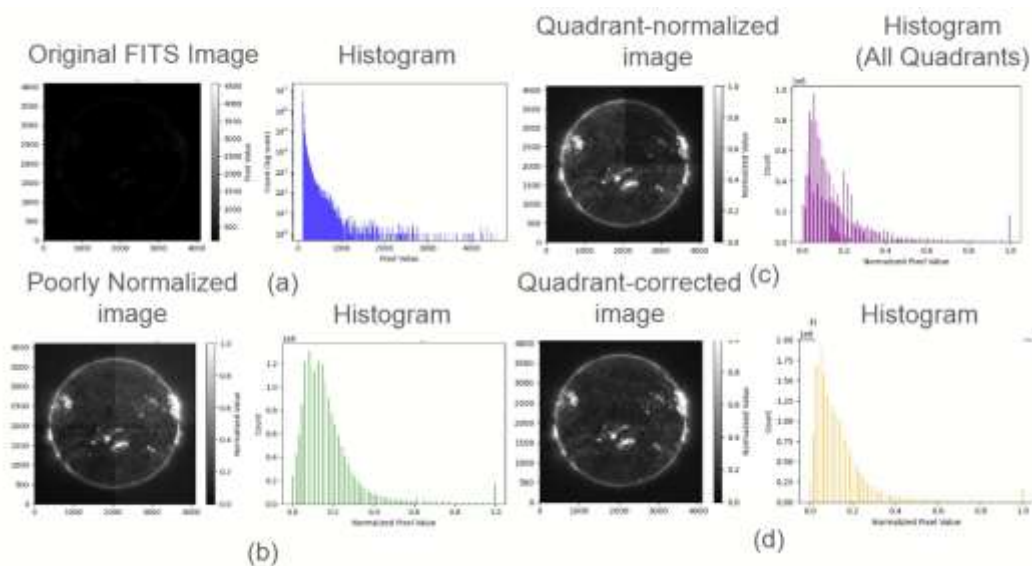


Figure 16: Illustrative examples of preprocessing a Level 0 SDO sample along the corresponding histogram: (a) Original raw sample, (b) Preprocessing based on the full-image, where all four quadrant discontinuities are clearly visible, (c) Per-quadrant preprocessing by clipping pixel values at the 1st and 99th percentiles, resulting in one remaining discontinuity (d) Per-quadrant preprocessing by shifting each quadrant's median to match the global median, followed by percentile based normalization.

Overall, we visually observe the algorithm outcomes as well as the IoU score, and we obtain scores similar to level-2 SDO samples. Interestingly, models trained on Level-2 data retain compatibility and perform satisfactorily on Level-0 observations. Moreover, we observe that the active regions models can be applied to even poorly normalized level-0 SDO, like illustrated in Figure 17.

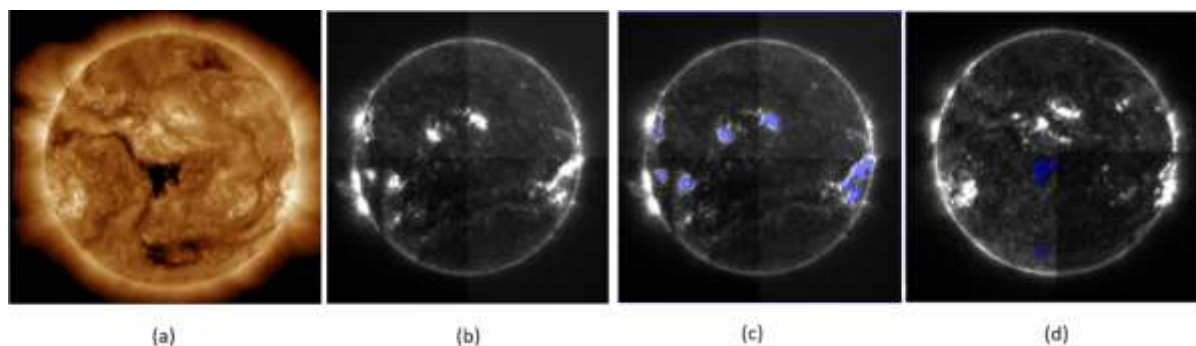


Figure 17: Evaluating SCSS-Net on Level 0 SDO AIA data using simpler than optimal preprocessing schemes: (a) Level 2 SDO AIA 193Å sample, (b) poorly preprocessed Level 0 SDO AIA 131Å sample (input), we can visually see the quadrant discontinuities (c) AR prediction overlay, (d) CH prediction overlay. AR SCSS-Net appears more tolerant than CH SCSS-Net when preprocessing is minimal and quadrant discontinuities are visible.

Through this study, we demonstrate that machine learning solutions developed for ground-based analysis can be adapted for autonomous operations in future solar missions. Our findings provide a foundation for designing frameworks tailored to onboard hardware configurations.

## 4. SCIENTIFIC RESULTS AND RECOMMENDATIONS

The results mainly consist of methods that can be re-used in further studies, and scientific knowledge (including publications) that increase our understanding of solar active regions and flares, with the goal of providing better forecast for space weather.

### 4.1 Flare predictions applying CNNs on line-of-sight magnetograms of individual active regions

#### 4.1.1 Evaluation of the classification of individual magnetograms

The result of the evaluations of our CNN models is shown in Table I. These values were calculated by averaging the result of the evaluation of each group of models across their cross-validation. At a first glance, the Trad-CNN models in the CMX classification section present equivalent results no matter the loss function, whereas the results of the SPP-CNN models indicate them to be better at predicting flares when trained using the BCE loss function on the CMX classification. The results indicate an improvement of 0.25 in PR AUC, 0.23 in TSS and 0.25 in precision. Moreover, the evaluation results of the SPP-CNN models trained using the BCE loss function are on average higher on every metric than the results of the Trad-CNN models trained with the TSS-based loss function. We notice in this case an improvement of 0.1 in the TSS metric, 0.17 in precision and a 0.11 improvement in PR AUC metric. Given the standard deviations of the TSS of our models, we cannot conclude one model to be strictly superior to another. However, using the Kolmogorov-Smirnov test on the TSS metric obtained throughout the cross-validation folds of our models, we acquire a confirmation at 92% chance for the SPP-CNN models trained with the BCE loss to follow a different probability distribution than SPP-CNN models trained with the SOL function. This allows us to infer that there is a 92% chance that using BCE to train SPP-CNN models was better than the TSS-based loss. On the other hand, the results of the evaluations of our models using the MX classification are more nuanced, there is no superior model. Furthermore, although the models achieve promising scores in recall and TSS, the best precision observed is 0.11 for the TSS-based loss trained SPP-CNN model, which is too low to consider using them for operational purposes or any explainable artificial intelligence (XAI) analysis.

Table I: Average results of evaluation of cross-validation of SPP-CNN and Trad-CNN based on classification and loss function.

Classification	Architecture	Loss	TP	TN	FP	FN	Accuracy	Precision	Recall	TSS	TSS Std <sup>(a)</sup>	PR AUC
CMX	Trad-CNN	BCE	1310	9584	2150	545	0.8	0.38	0.7	0.52	<u>0.02</u>	0.52
		TSS	1369	9713	2118	502	0.81	0.39	0.73	0.55	0.06	0.57
	SPP-CNN	BCE	1401	10429	1253	464	<u>0.87</u>	<u>0.56</u>	<u>0.76</u>	<u>0.65</u>	0.16	<u>0.68</u>
		TSS	1374	7834	3691	497	0.69	0.31	0.74	0.42	0.06	0.43
MX	Trad-CNN	BCE	242	7622	5779	71	0.58	0.06	0.77	0.35	0.17	0.08
		TSS	177	10884	2323	93	0.82	0.08	0.67	<u>0.5</u>	<u>0.09</u>	0.11
	SPP-CNN	BCE	174	10827	1994	107	0.84	0.1	0.62	0.46	0.24	0.14
		TSS	217	9460	3614	81	0.73	<u>0.11</u>	0.7	0.43	0.28	<u>0.2</u>

Notes. <sup>(a)</sup>True skill statistic standard deviation. Underlined values highlight the best result between every models using the same classification.

While these statistics and indices highlight the overall prediction performance of our models, they do not give any information on the type of flare our models are able to predict. Using a CMX classification, our model could have a recall of 0.8 while correctly predicting only C class flares. To acquire a better understanding of the capability of our models, we analyzed their predictions results on the test dataset for each image label and the evolution of the prediction confidence of a flare, i.e. the output value of the last neuron of our models, during the entire lifetime of a given active region.

Table II and Table III present the number of correct predictions, the accuracy and the standard deviation of the accuracy, based on the label, the model, the loss function and the classification. Figure 18 and Figure 19 illustrate the best, the worst, and the average prediction accuracy per label for every model trained. These tables and figures indicate an improvement in prediction performance of SPP-CNN models compared to Trad-CNN models for the CMX classification. The first result we can note is the perfect prediction accuracy of the SPP-CNN models for images linked to X class flares and very high accuracy for images linked to M class flares. The second result is the very low accuracy for FX-labeled images compared to other images.

Table II: Label-wise average results of SPP-CNN and Trad-CNN evaluations using the CMX classification.

Classification	Model	Loss	Class	Label	Correctly Predicted	Wrongly Predicted	Acc <sup>(a)</sup>	Acc Std <sup>b</sup>
CMX	Trad-CNN	BCE	1	X	23	4	0.83	0.33
				M	222	18	0.92	0.03
				C	1064	523	0.67	0.04
			0	NF	7840	1020	0.89	0.02
				FX	19	14	0.36	0.41
				FM	128	140	0.48	0.2
		TSS	1	X	18	4	0.82	0.37
				M	238	19	0.92	0.04
				C	1113	478	0.69	0.08
			0	NF	7883	1067	0.88	0.02
				FX	41	34	<u>0.68</u>	0.29
				FM	1621	893	0.64	0.06
	SPP-CNN	BCE	1	X	26	0	<u>1.0</u>	0.0
				M	291	10	<u>0.97</u>	0.04
				C	1084	453	<u>0.71</u>	0.15
			0	NF	8449	589	<u>0.93</u>	0.06
				FX	17	2	<u>0.5</u>	0.5
				FM	185	75	<u>0.76</u>	0.15
		TSS	1	X	33	0	1.0	0.0
				M	228	23	0.9	0.06
				C	1112	473	0.7	0.12
			0	NF	6236	2446	0.72	0.12
				FX	7	30	0.18	0.37
				FM	107	190	0.37	0.11
FC	1483	1023	0.59	0.17				

<sup>(a)</sup> Accuracy <sup>(b)</sup> Accuracy Standard deviation

Notes. Value underlined highlight the best result among all models and Loss function.

Table III: Same as Table II but using the MX classification.

Classification	Model	Loss	Class	Label	Correctly Predicted	Wrongly Predicted	Acc <sup>a</sup>	Acc Std <sup>b</sup>	
MX	Trad-CNN	BCE	1	X	19	7	0.73	0.38	
				M	158	86	0.66	0.13	
			0	C	806	791	0.5	0.06	
				NF	7811	885	0.9	0.09	
				FX	12	11	<u>0.52</u>	0.48	
				FM	219	70	<u>0.75</u>	0.16	
				FC	2035	564	0.78	0.06	
			TSS	1	X	15	11	0.71	0.35
					M	227	60	<u>0.79</u>	0.12
	0	C		564	920	0.37	0.22		
		NF		5563	3445	0.63	0.27		
		FX		28	41	0.44	0.19		
		FM		135	89	0.51	0.28		
		FC		1330	1281	0.51	0.24		
	SPP-CNN	BCE		1	X	21	5	<u>0.78</u>	0.39
					M	195	75	0.69	0.39
			0	C	853	695	0.53	0.3	
				NF	6780	1948	0.79	0.23	
FX				28	5	0.24	0.42		
FM				189	80	0.71	0.27		
FC				1608	883	0.65	0.3		
TSS			1	X	19	7	0.6	0.49	
				M	154	99	0.61	0.32	
	0	C	950	542	<u>0.64</u>	0.21			
		NF	7671	772	<u>0.91</u>	0.07			
		FX	15	6	0.41	0.43			
		FM	271	137	0.68	0.28			
		FC	1918	535	<u>0.79</u>	0.16			

<sup>(a)</sup> Accuracy <sup>(b)</sup> Accuracy Standard deviation

Notes. Value underlined highlight the best result among all models and Loss function.

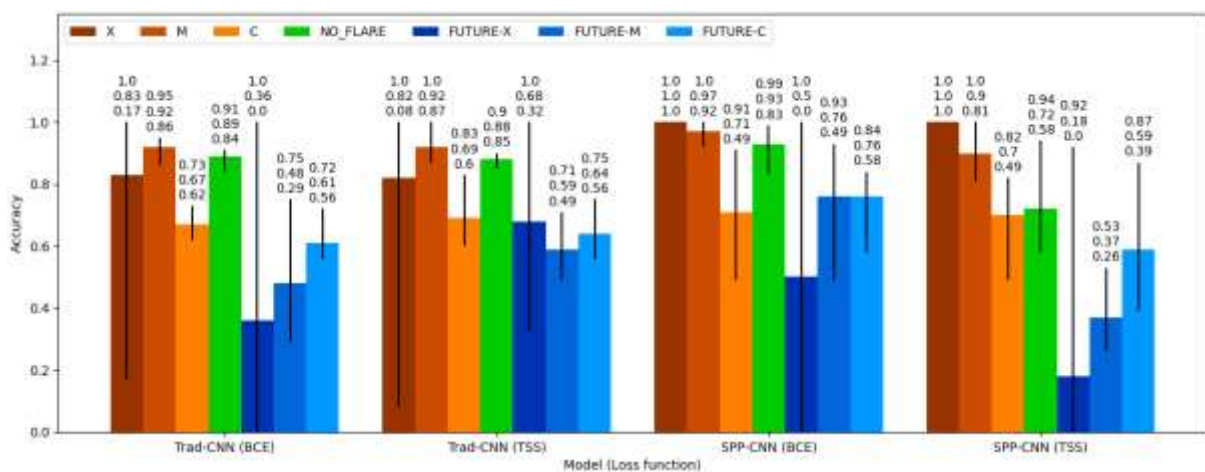


Figure 18: Prediction accuracy of models trained on the CMX classification, as distributed between each label. The error bars on each bar of the chart represent the range of accuracy between the best and worst model for each label. The three numbers above each bar represents, from top to bottom, the maximum accuracy, the average accuracy, and the minimum accuracy for each label.

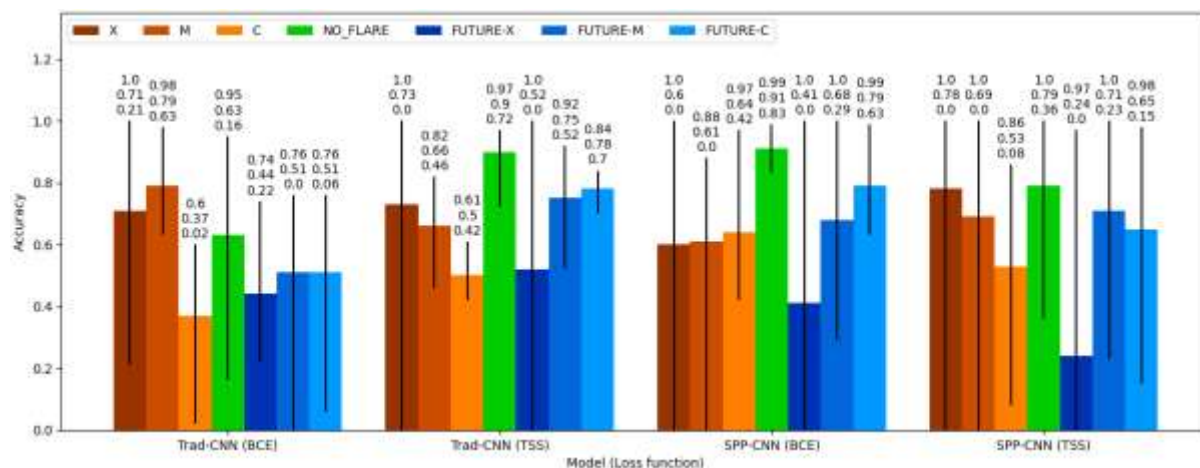


Figure 19: Same as Figure 18 but for the MX classification.

#### 4.1.2 Monitoring of the time variation of the classification

To monitor how the classification by the models evolve as the active region evolve, with respect to the time delay before the next flare, Figure 20 to Figure 27 highlight the evolution of the prediction confidence of the models trained using the CMX classification, for the prediction of every image of a single active region. Figure 20 to Figure 23 present the evolution of the prediction confidence when predicting images from ARs producing X class flares. Meanwhile, Figure 24 to Figure 27 present the evolution of the prediction confidence for images from ARs producing an M class flare as the strongest flare of their lifetime. Starting from the first image of each active region within the test dataset, the x-axis represents the time in hours, with markers illustrating the occurrences of solar flares produced by the corresponding active region. The y-axis represents the prediction confidence of the corresponding model, and the horizontal line at 0.5 represents the threshold that separates a positive prediction (1) from a negative prediction (0). To ease the analysis of the figures, parts of the chart with the colored area correspond to the interval of time during which images are labeled positive in our test dataset.

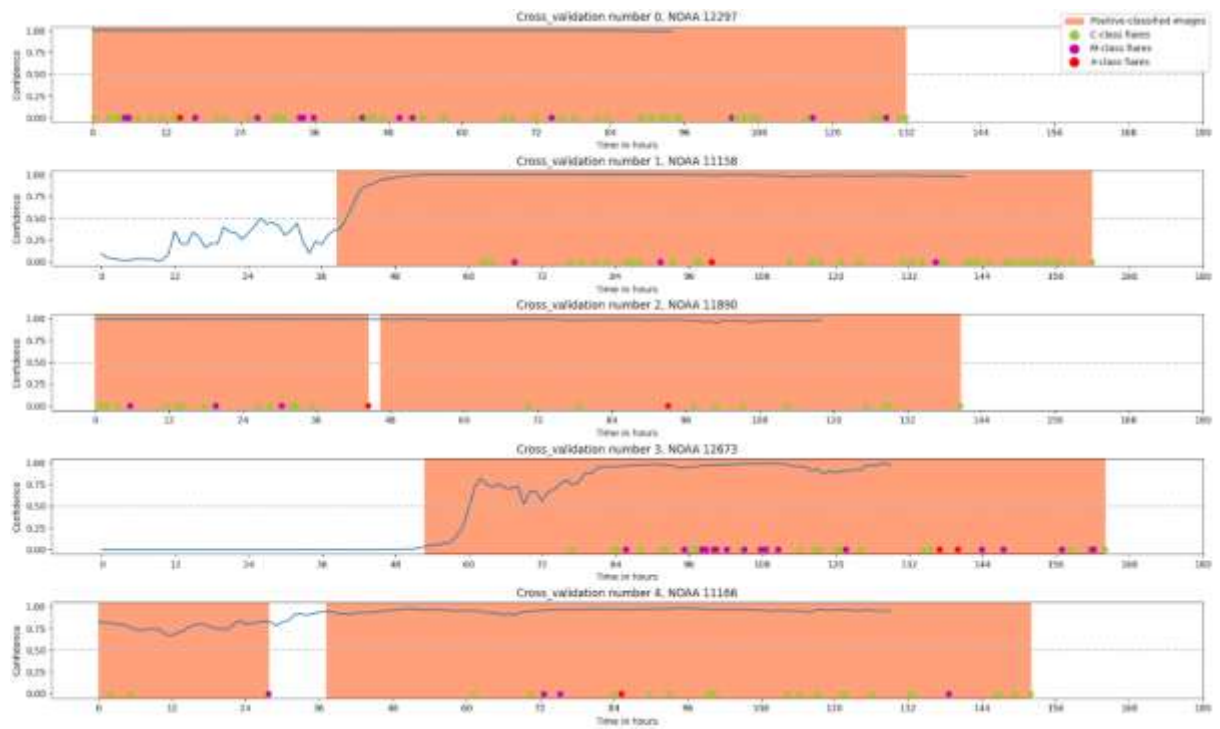


Figure 20: Prediction confidence for the prediction of images of active regions producing X class flares by SPP-CNN, trained with BCE loss using the CMX classification. The colored area corresponds to the interval of time during which images are labeled positive in our test dataset.

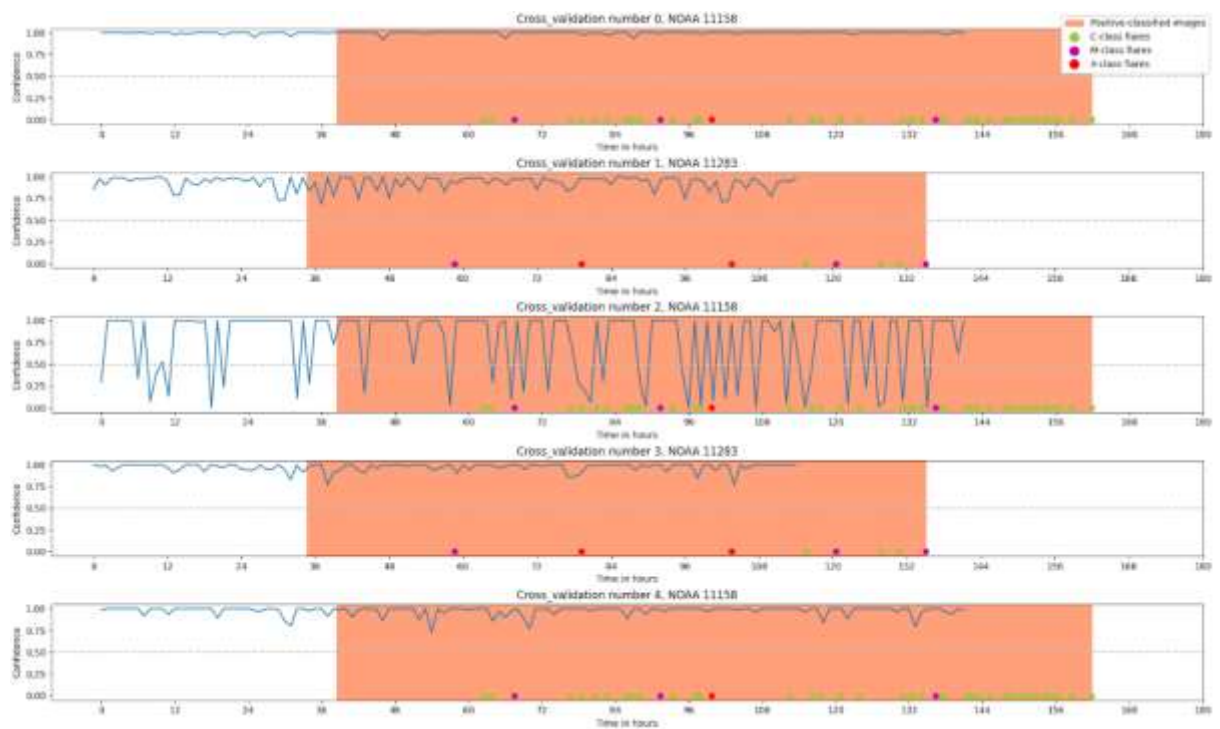


Figure 21: Same but with models using the SPP-CNN architecture and the SOL function.

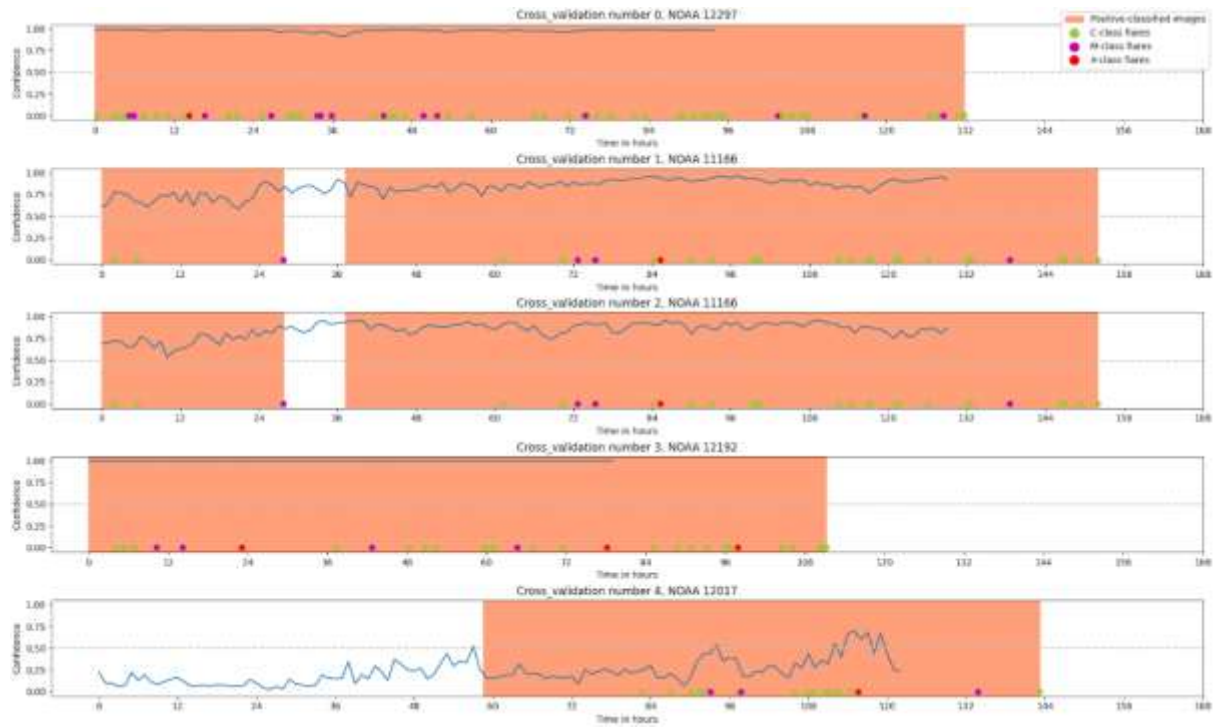


Figure 22: Same but with models using the TRAD-CNN architecture and the BCE loss function.

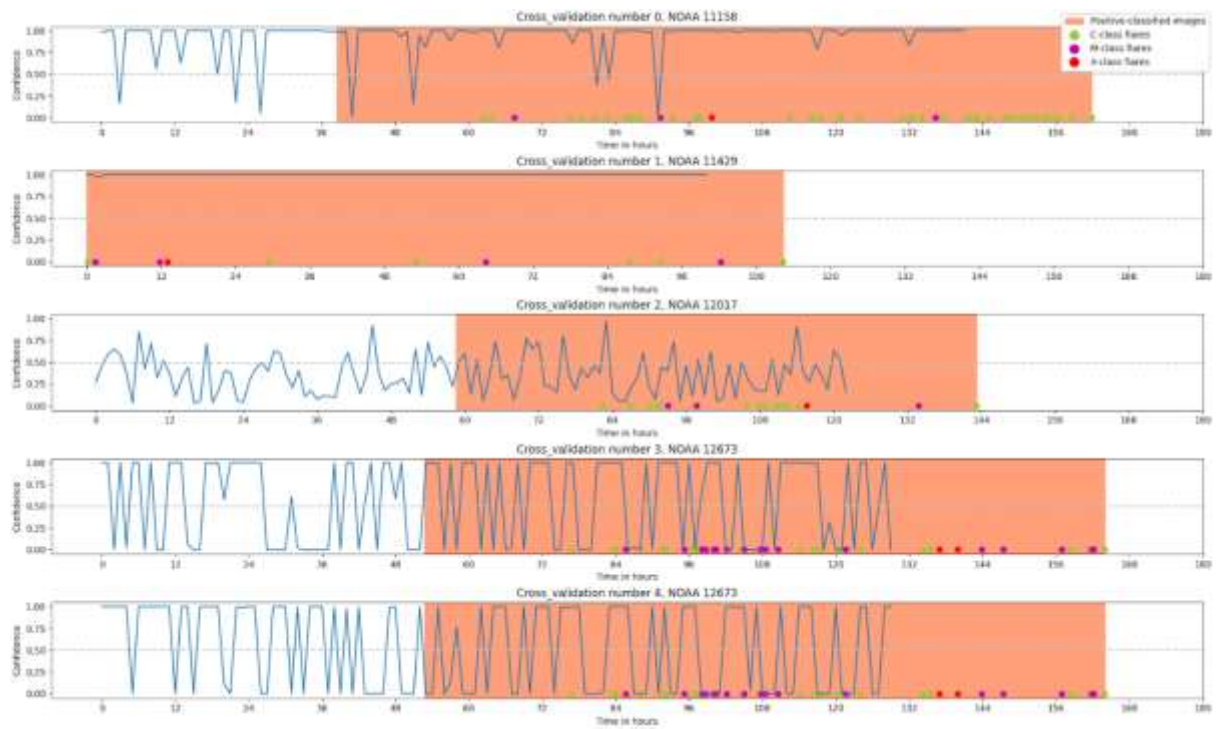


Figure 23: Same but with models using the TRAD-CNN architecture and the SOL function.

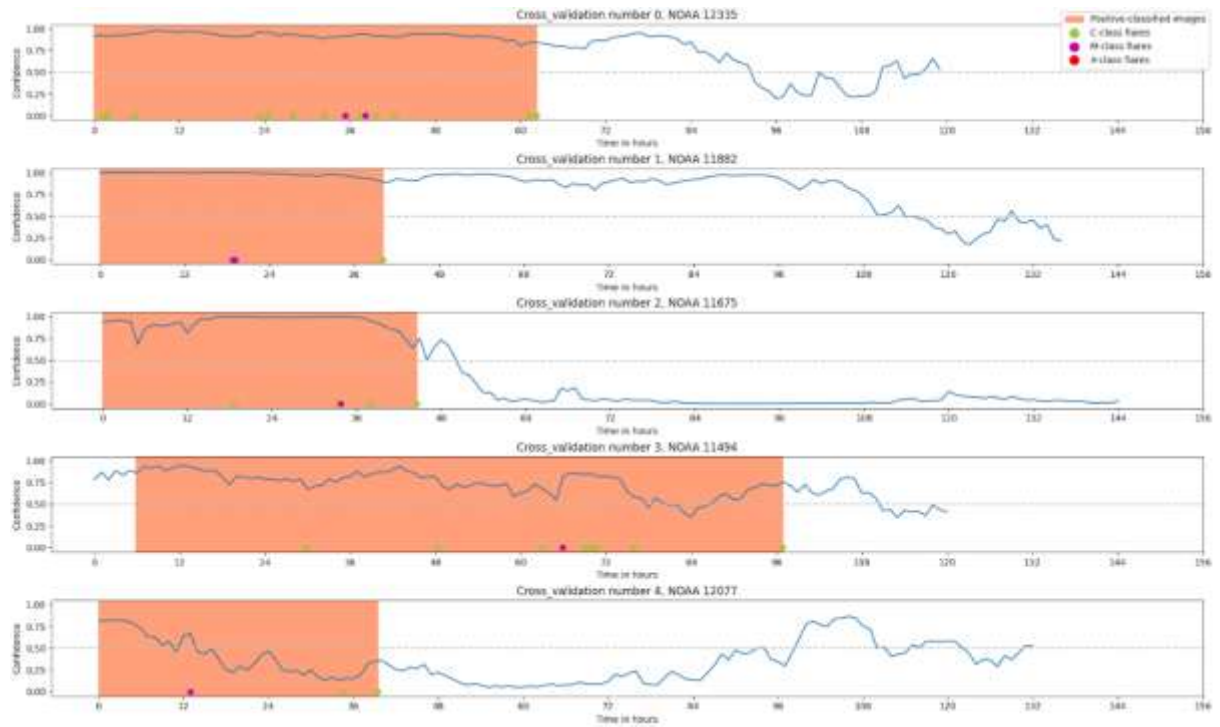


Figure 24: Prediction confidence for the prediction of images of active regions producing M class flares as the strongest flare by SPP-CNN, trained with BCE loss using the CMX classification. The colored area corresponds to the interval of time during which images are labeled positive in our test dataset.

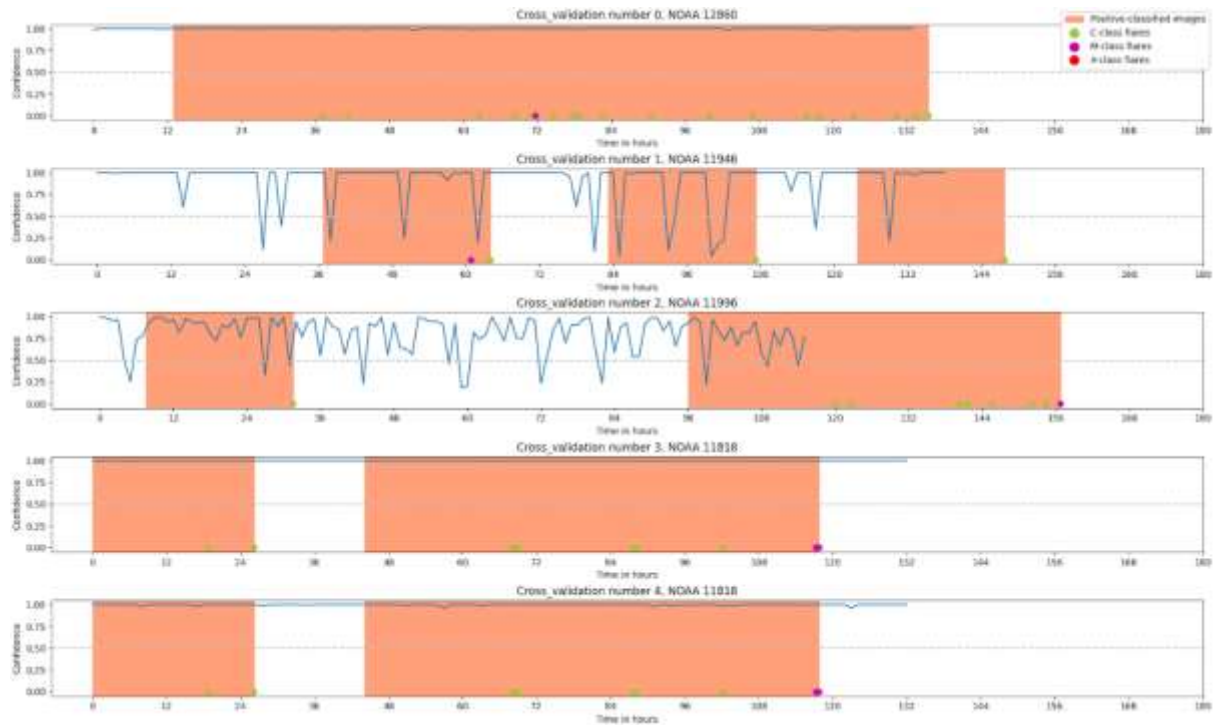


Figure 25: Same but with models using the SPP-CNN architecture and the SOL function.

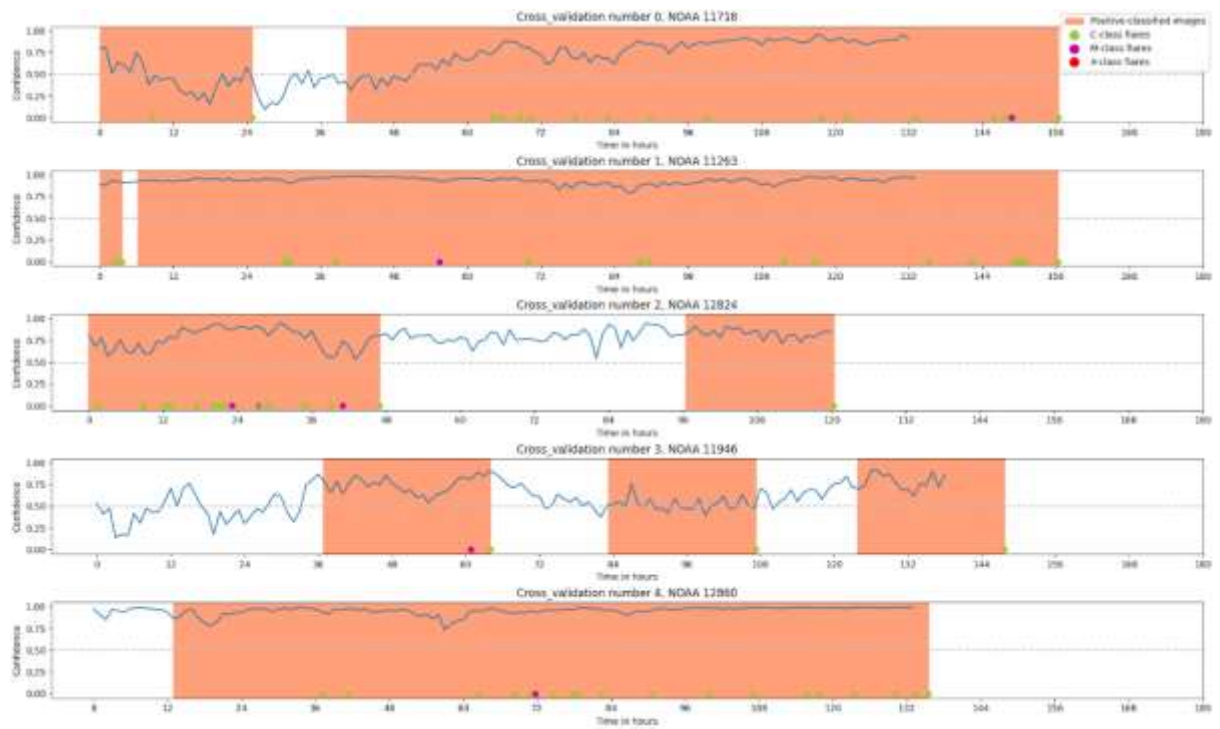


Figure 26: Same but with models using the TRAD-CNN architecture and the BCE loss function.

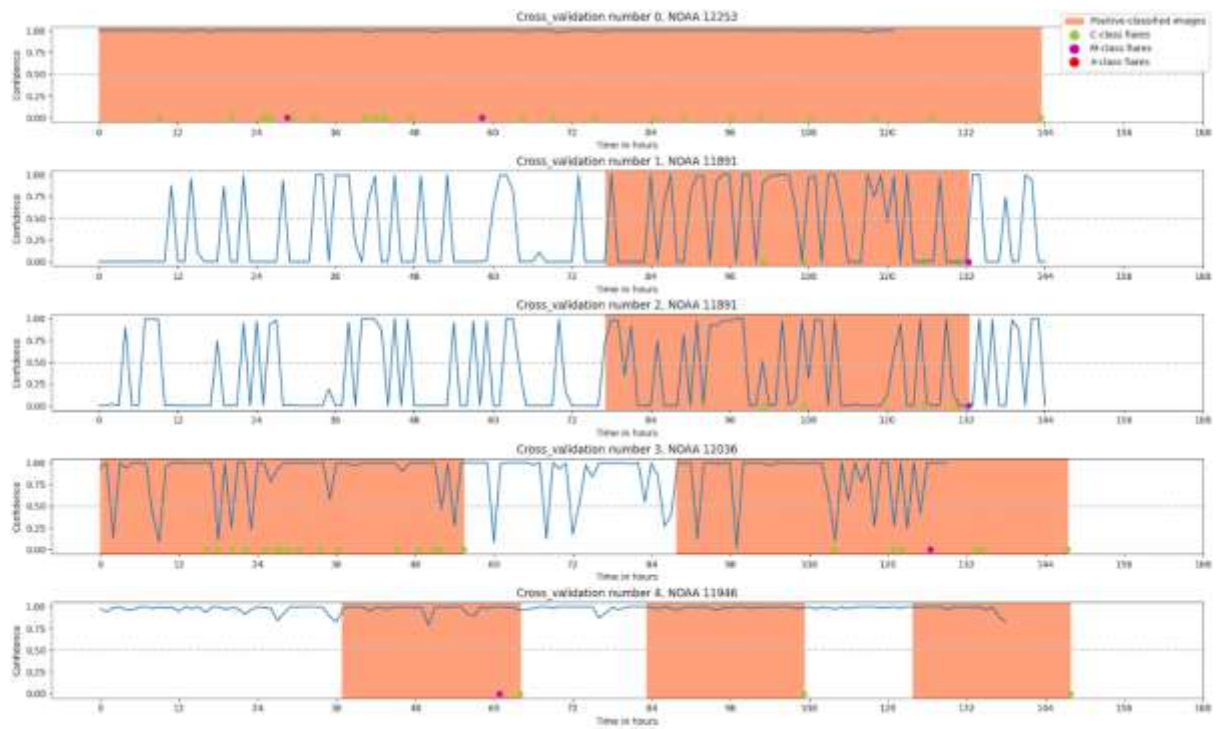


Figure 27: Same but with models using the TRAD-CNN architecture and the SOL function.

#### 4.1.3 Comparison of the results between architectures and loss functions

The first noteworthy point in our results is the relative improvement of the SPP-CNN compared to the Trad-CNN to predict flares using the CMX classification. There is indication that associating the SPP layer with the BCE loss function increase the rate of successful prediction of flares while reducing the rate of false alarm compared to every other couple of architectures and loss functions. An average increase of 0.1 in the TSS metric compared to Trad-CNN models using the TSS-based loss and a perfect classification of X-labeled images is especially notable.

While SPP-CNN models trained with the BCE loss function present impressive results, we can observe a drop in our metrics when using the Score-Oriented-Loss function instead of the BCE loss function. We speculate the SPP layer to be incompatible with the more deterministic predictions offered by the SOL function compared to the BCE loss function, which can be observed in Figure 20 to Figure 27. The process of SPP consists in slicing its input (i.e, feature maps) in tiles and selecting the maximum value in each of those tiles. This means extracting the presence and intensity of a feature within a tile while factoring out the quantity of such feature. This process transforms the output of each tile to either “presence” or “non-presence” of the feature along with its intensity and can overlap with the deterministic prediction associated with the SOL function.

While the prediction ability of SPP-CNN models decrease when using a SOL function, Trad-CNN models present a slight improvement in prediction ability when trained with it compared to the BCE loss function. Although the gap of TSS is more noticeable for the MX classification, the difference is so small we cannot deduce whether the use of a SOL is the source of such improvement or if it originates from the random splits of our dataset. We would need further cross-validation fold to obtain a reliable average. Nevertheless, our results show potential in the use of a SOL function for flare prediction using more traditional architectures of models, although the prediction confidences in Figure 24 to Figure 27 look more uncertain through the numerous sudden swap in prediction during the evolution of a single active region.

Aside from the difference linked to the impact of the loss function between our architectures, we can observe the stability of the training of our SPP-CNN models in Figure 18. The classification accuracy of images of active regions leading to stronger flares is especially stable compared to active regions producing weaker flares or not producing any flare at all. The accuracy of the SPP-CNN models also decreases with the strength of observed flares whereas Trad-CNN models classify more accurately images of active regions producing flares of medium to low intensity or not producing any flare at all compared to strong, X flares.

#### 4.1.4 Comparison of the results between classifications systems

The second point to note is the disparity in recall, TSS and especially precision and PR AUC throughout the results of our models between the CMX and MX classifications, no matter the loss function. The average (Recall, TSS, Precision, PR AUC) values for all models using the CMX classification are (0.73,0.53,0.41,0.55) while they become (0.69,0.43,0.09,0.13) for models using the MX classification.

The results for the CMX classification demonstrate the excellent ability of our models to predict flares in the next 24 hours, especially with the SPP-CNN architecture. However, they are unable to output any reliable positive prediction using the MX classification. The average precision of 0.09 highlight the inability to correctly differentiate a negatively labelled image from a positive one. Furthermore, the disparity of prediction ability of our models between the CMX and the MX classifications are clearly

visible while comparing Figure 18 and Figure 19. The accuracy and error bars of Figure 19 show an instability of prediction performance among the models of a same classification and loss function, and their inability to correctly classify images between positive (flare) and negative (non-flare). It would be interesting to obtain a deeper view of the functioning of our models and where their attention lies.

The gap between the results of our models trained on the CMX classification and the MX classification can be explained by two reasons: the extreme class imbalance in the datasets of the MX classification and the increasing difficulty in differentiating flares for the MX classification due to the presence of C-labelled images in the negative class.

Firstly, including C-labelled images in the negative class lead us to a dataset distribution of 2% of positive-classified images to 98% of negative-classified images (to be compared to 13% and 87%, respectively, in the case of the CMX classification). Although several efforts were made to reduce the class imbalance in this study such as using class-weights, downsampling and Score-Oriented-Loss function, the number of positive-class images is too low (2000 images used for training and 200 for the test dataset) to properly train our models.

Secondly, the presence of C-labelled images in the negative class raises three major problems:

1. The first problem is linked to the presence of C class flares in active regions producing M or X flares. While the triggering mechanisms of a C class flare may be different from an M or X flare, an active region producing a C flare followed by a higher class flare a few hours later may not present a significant evolution enough to be seen in a line-of-sight magnetograms. The similarity of images from such active regions can confuse our models in the way that two magnetograms observed within a 1-hour interval are supposed to be classified as positive and negative, respectively, without significant differences.
2. The second problem raised is linked to the classification used in the GOES catalogue. The five classes of the GOES classification (A, B, C, M and X) follow a logarithmic scale representing the intensity of soft X-ray emission of a flare and each flare class is composed of subclasses from 1 to 9. This is useful when discussing flares, but when it comes to training a CNN, it becomes harmful to our prediction performance. This is because, in the GOES classification, a C9.0 flare will be classified in another class than an M1.0 flare while not presenting a significant difference. This is also the case between M9.0 and X1.0 flares and between B9.0 (which is labelled as an NF image in our labelling system) and C1.0 flares. Labelling our images based on this classification model can be confusing for the DL model as two images without significant differences can be in separate GOES classes and therefore in separate data classes. Furthermore, this issue can explain the ability of our models to correctly predict stronger flares better than weaker ones in the positive class. It is harder to differentiate images of flares close to the threshold separating flaring and non-flaring images.
3. The third problem can explain the difference of prediction performance between CMX and MX classifications for models following the SPP-CNN architecture: some features leading to solar flares may appear in an active region with an incoming flare of class C, M, and X flares. The SPP can efficiently detect the presence and intensity of such features. However, one of the differences between C, M, and X-labelled line-of-sight magnetograms may be its quantity. While they may be distinguishable in a regular CNN, it is hardly possible with SPP-CNN models.

As mentioned before, this is because the SPP layer slices its input into a fixed amount of tiles before selecting the maximum value in each tile, extracting the presence and intensity of a feature while factoring out the amount of time it appeared. If we were to input an X-labelled image with a very complex structure and a feature that appears multiple times, it would produce the same output as if the feature had appeared only once in a simple structure from a C-labelled image. This can be interpreted as if our SPP-CNN models can process features inherent to the triggering of solar flares, no matter their GOES class. However, their prediction abilities may be solely based on these features and whenever we mix C-labelled images in the negative class, it becomes ineffectual. In other words, the models may be able to predict flares with precision, but they do not differentiate between their class at all. This is the drawback of the coarse slicing used in the SPP layer and can be bypassed by increasing its parameters, which represents the number of tiles the input has to be sliced into. However, this would also drastically increase the computational power needed to train our models and to run a prediction.

The drop of predictive performance from models trained on the CMX classification to models trained on the MX classification is also observed in other studies, such as Li et al. (2020).

#### **4.1.5 Analysis of results label-wise**

A third noteworthy point is the significantly lower prediction accuracy of FX-labeled images compared to all other labels. This may be due to the large range of time intervals between the manifestation of flare triggering mechanisms and the actual flare. An active region may keep a complex magnetic configuration during several days before triggering a flare, trigger a flare a few hours after the manifestation of flare triggering mechanisms, or present important signs of an incoming eruption while not erupting at all. We can observe the impact of this phenomenon in Figure 20 to Figure 27, especially in the chart of the evolution of the prediction confidence of the "Cross\_validation number 1" in Figure 20 where the prediction confidence remains close to 1 during several days after the last eruption and in the "Cross\_validation number 3" chart of Figure 24 where the prediction confidence increase only 16 hours before the first flare. In these conditions, labelling our data based on a fixed prediction window separating images of flaring (positive) and non-flaring (negative) active regions can greatly confuse our models and impede on their prediction ability.

#### **4.1.6 Effect of image scaling processes of the data on the prediction performances**

Figure 28 shows the evolution of the recall, precision and TSS metrics throughout the variation of the amplitude of the image scaling processes. The horizontal dashed lines highlight the average values of metrics of the models using the original dataset. The vertical dashed line points to the ratio for which no image scaling process was used on the dataset. The values showcased on this plot are the average value over five-fold cross validation.

We can first notice a drop in recall and TSS for both of those plots whenever the ratios are diverging from 1. This means the positive predictions as well as the overall predictive abilities of the models decrease in quality no matter which image scaling process and no matter which amplitude is applied. However, there is a significant drop in those metrics when the ratios fall under 0.5, especially compared to cases when ratios are higher than 1.

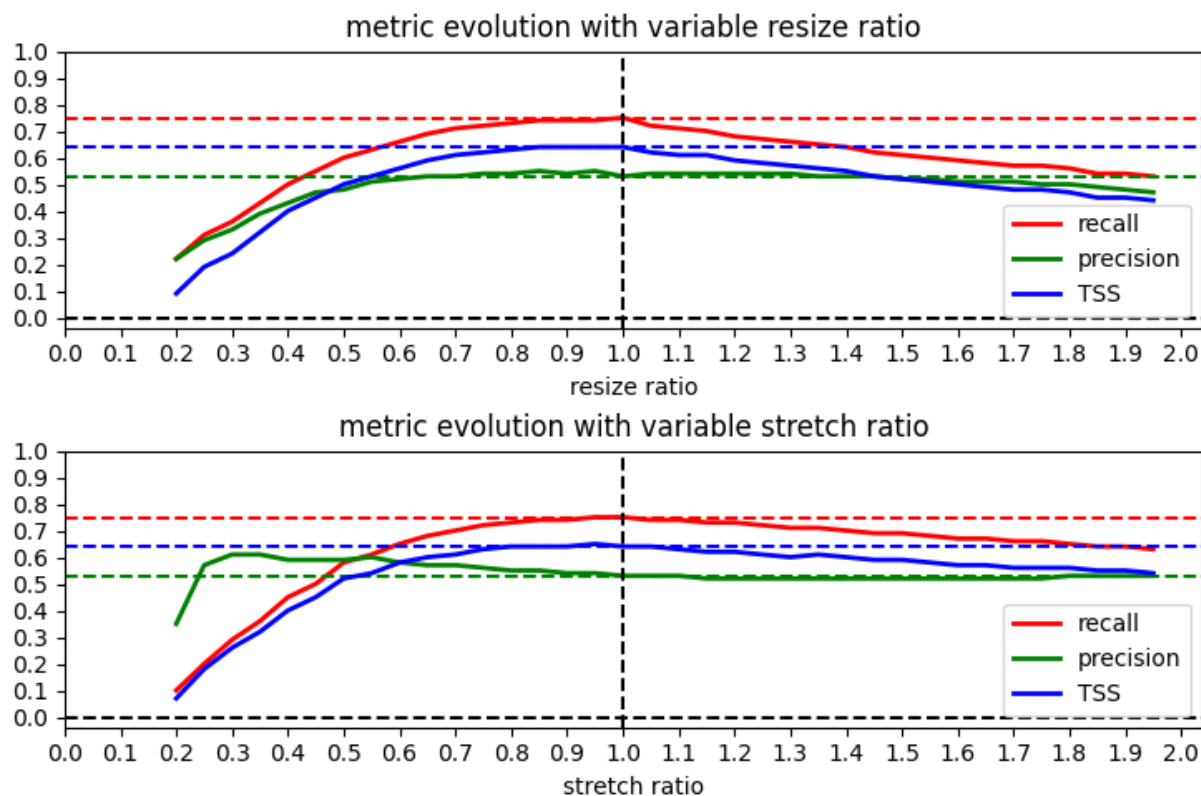


Figure 28: Evolution of the metrics of the prediction ability of SPP-CNN models following the application of image scaling processes on the test sub-datasets.

Although the recall and TSS metrics decrease due to image scaling processes, the precision only decreases when the models are faced with the resizing process. The stretch process only negatively impacts the precision of the models when the ratio is 0.2. Otherwise, the precision stays near the original value or rises when the stretch ratio decreases. This may be due to the models outputting mainly negative predictions in those cases.

There is a correlation between the size of the image and the intensity of a future flare. Images of an AR that will produce a flare of higher intensity tend to be of higher size. This leads us to wonder about the difference of resizing and stretching amplitude for different input sizes. Table IV showcases the stretch ratio and in turn the TSS variation caused by the stretching effect when the input size of the models are square (i.e., when the width of the input is the same as the height). We notice a loss of 0.1 TSS for M, C and NF sub-labelled images, whereas the stretching of X sub-labelled images is linked to a loss of only 0.04 TSS. The impact of the stretching process when using a square input size on the prediction performance of a deep learning is relatively low and may be negligible depending on the use of the models.

Table IV: Average stretch ratio and TSS variation per sub-label for squared input size.

Value	X	M	C	NF
Stretch ratio	0.64	0.54	0.54	0.57
TSS variation	-0.04	-0.1	-0.1	-0.1

However, Table V showcases the average ratio of resizing for each label for a given square size, which is usually used in deep-learning methods, as well as the linked variation in TSS of the models. The square sizes presented in this table vary from 100×100 to 500×500, with increments of 50×50. We can notice the resize ratios 450×450 and 500×500 link to a relatively low loss of TSS for positively labelled images and a notable loss of TSS for negatively labelled images. On the other hand, resizing images to a lower square size is linked to significant loss of predictive ability in our models. Square sizes lower than 300×300 induce a resize ratio lower than 0.25, which is linked to a TSS variation of less than -0.45.

Table V: Average resize ratio and TSS variation per sub-label for different square size.

Value	Sub-label	100×100	150×150	200×200	250×250	300×300	350×350	400×400	450×450	500×500
Resize ratio	X	0.03	0.06	0.11	0.18	0.25	0.34	0.45	0.57	0.7
	M	0.03	0.07	0.12	0.18	0.26	0.36	0.47	0.59	0.73
	C	0.04	0.09	0.16	0.25	0.36	0.49	0.65	0.82	1.01
	NF	0.09	0.21	0.38	0.59	0.84	1.15	1.5	1.9	2.34
TSS variation	X	-0.55	-0.55	-0.55	-0.55	-0.45	-0.32	-0.19	-0.11	-0.03
	M	-0.55	-0.55	-0.55	-0.55	-0.45	-0.32	-0.19	-0.08	-0.02
	C	-0.55	-0.55	-0.55	-0.45	-0.32	-0.14	-0.05	-0.01	0.0
	NF	-0.55	-0.55	-0.24	-0.08	0.0	-0.03	-0.12	-0.19	-0.2

Furthermore, in this study, we scaled the test sub-datasets to a minimum ratio of 0.2 for both resizing and stretching processes. For each case where the resize ratio was lower than 0.2 in Table V, the TSS variation was set as if using a resize ratio of 0.2. This means that the TSS variation of X and M sub-labeled images may be lower than what is shown in the Table for square sizes between 100×100 and 250×250. While using a square size lower than 350×350 incur a non-negligible TSS variation, the use of a square size of 500×500 only incurs a loss of TSS of 0.2 for NF sub-labeled images and negligible TSS variation of other images.

These results suggest the image scaling processes can greatly impact the prediction ability of SPP-CNN models. We notice a drastic decrease of TSS when resizing or stretching images with a ratio of over 1.4 or lower than 0.7. Although the loss of prediction ability can be negligible with a ratio between 0.7 and 1.4, this is a case rarely used for deep-learning models in flare forecasting. The input size used for deep-learning models in flare forecasting is usually between 100×100 and 250×250. Furthermore, as there is a relation between the image size and the complexity of an AR, which may also be related to the intensity of a flare, images of ARs that will produce flares of high intensity are impacted to a higher degree by the image scaling processes.

#### 4.1.7 Regions of attention of the Neural Network

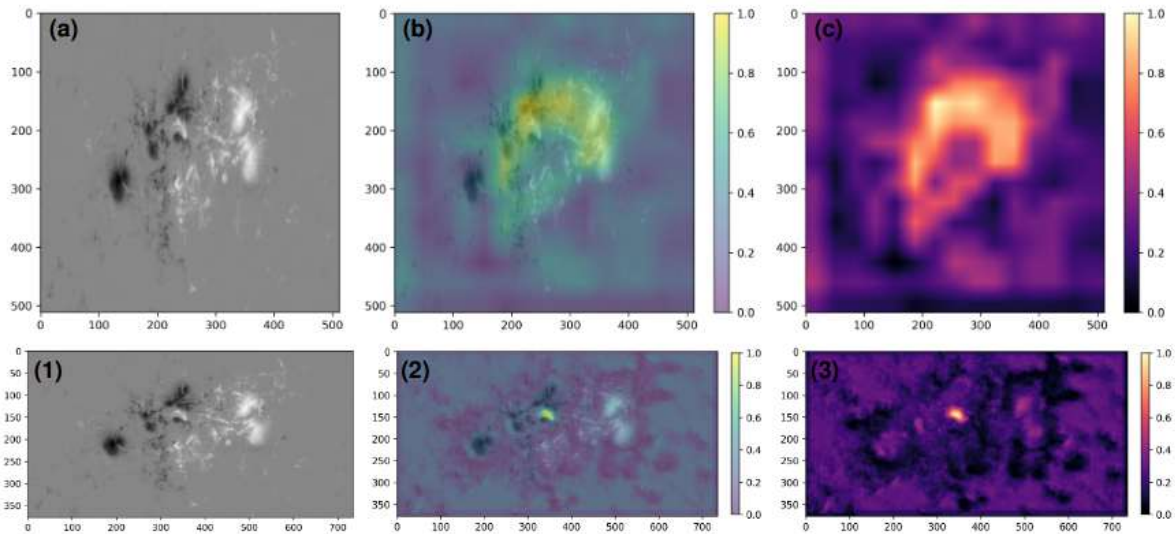


Figure 29: Images illustrating the difference of attention between the Trad-CNN and SPP-CNN models, using Grad-CAM. (a) is the original image resized for Trad-CNN, (b) is the superposition of (a) and (c) and (c) is the heatmap produced using Grad-CAM on the classification of (a) using Trad-CNN. (1) is the original image used by SPP-CNN, (2) is the superposition of (1) and (3) and (3) is the heatmap produced using Grad-CAM on the classification of (1) using SPP-CNN.

It is interesting to visualize the difference of vision between a traditional model, which uses resized, squared images, and a model using the SPP layer. This allows it to bypass image scaling processes before its input, and in turn also allows the model to use non-scaled images. To achieve this goal, we used the Grad-CAM method to highlight regions of the image that contributed the most to the predictions of the models.

Figure 29 showcases the output of the Grad-CAM method through six images spread into two rows of three columns. On one hand, the first row of images is linked to Trad-CNN, while the second row of images is linked to SPP-CNN. On the second hand, the three columns each correspond to a type of image:

- The first column corresponds to the original image used by the models for their prediction. The NOAA number of the AR highlighted, which is an identification number attributed by the National Oceanic and Atmospheric Administration (NOAA), is 11166. Furthermore, the image was taken on October 9th of 2011 at 08:00, preceding an X1.5-class flare at 23:13 the same day.
- The third column corresponds to the output of the Grad-CAM method. This image is a heatmap, highlighting the important parts of the image through bright colours.
- The second column showcases the superposition of the image used by the model for the prediction (column 1) and the heatmap (column 3).

The heatmap of the attention of the Trad-CNN model showcases attention-regions of notably larger size than the heatmap of the attention of the SPP-CNN model. The bright parts of the heatmap of Trad-CNN also cover a large part of the image. However, the heatmap of SPP-CNN highlights with high

indication of attention a precise region of the image. The highlighted part corresponds to a region importantly surrounded by its inverse polarity. Although there is a noteworthy difference in size of the bright patches between the heatmaps of SPP-CNN and Trad-CNN, both models also seem to use parts of the image which show no activity from the AR.

Several studies previously already related the flare potential of an AR to the length of Polarity Inversion Lines (PILs; Jing et al., 2006). Furthermore, researchers also highlighted the importance of the PIL of an AR for the flare prediction process of deep-learning methods using LOS magnetograms (Huang et al., 2018). This study confirmed this result, and highlighted how the SPP-CNN model better catch this feature.

During this study, we also tried to quantify the relation between the attention of SPP-CNN models and the PILs in order to obtain concrete proof of the relation. To do this, we first masked out the pixels that have a value under a given threshold in the heatmap output by Grad-CAM. After this, we calculated the average distance between each remaining pixel and their nearest pixel belonging to a PIL.

A pixel belonging to a PIL can be defined by calculating its partial derivatives and filtering the noise using:

$$1) f(x-2,y) * f(x+2,y) < T_{\text{mask}}$$

and

$$2) f(x,y-2) * f(x,y+2) < T_{\text{mask}}$$

In these equations  $f(x,y)$  is the value of the pixel at the coordinates  $x$  and  $y$  of the image, and  $T_{\text{mask}}$  a given threshold. In this study, we use  $T_{\text{mask}} = -0.5$ .

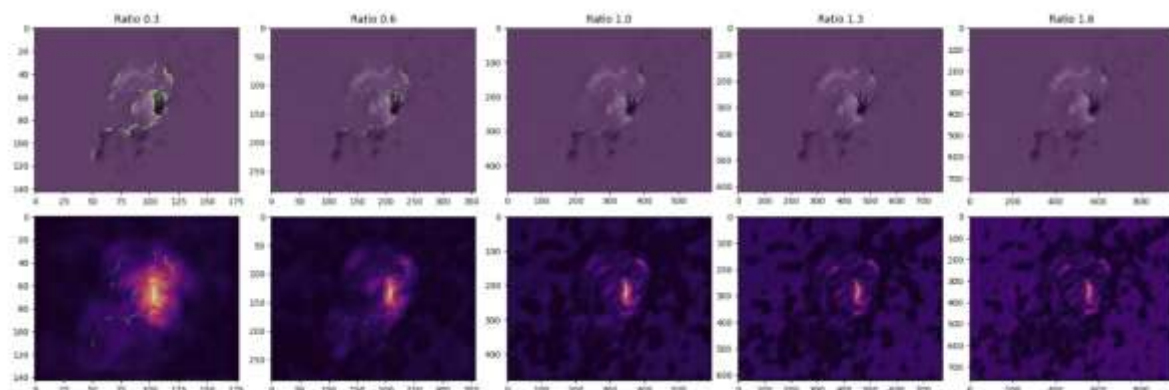


Figure 30: Highlight of the PILs in resized images with resize ratio of 0.3, 0.6, 1.0, 1.3, and 1.6, and the outcome of the Grad-CAM using those images with SPP-CNN.

Figure 30 shows the PILs detected using this method by superposing the map of pixels belonging to PILs and the original image. Table VI presents the average distance between a pixel in the attention-region of the models to its nearest PIL. In this Table, we can notice that the average distance between a pixel and a PIL decreases in inverse proportionality to the intensity of attention of the model. Regions

with high attention present pixels close to PILs contrary to regions with lower attention. However, this is not the case when using the image resized to a ratio of 0.3. In this case, pixels in the regions of attention of the models have a low average distance from the nearest PIL, no matter which  $T_{\text{heatmap}}$  used. We suppose that due to the high compression rate of the image, most of the pixels of the scaled images are close to a PIL.

Table VI: Average distance between pixels with Grad-CAM values above threshold  $T_{\text{heatmap}}$  and the nearest PIL.

Ratio	Average distance		
	T=0.5	T=0.7	T=0.9
0.3	1.45	1.54	1.22
0.6	2.18	1.83	1.61
1.0	2.92	2.10	1.82
1.3	6.73	2.02	1.36
1.6	8.24	2.15	1.34

In order to confirm the assumption that the attention of the model is focused solely on the PILs, we also plotted the histograms of the distance between every pixel in an attention-range and their nearest PIL to verify the unimodality of the distribution in Figure 31. We can notice an unimodality in most of the subplots and faint bimodalities in few sub-figures. The bimodalities in the sub-figures may be inconsequential due to the low differences in heights of the second peaks and the baselines of the distributions.

As the main peak of the sub-plots tend toward a distance of 1 between the pixels in high attention of the models and their nearest PIL, Figure 31 supports the importance of PILs for the prediction of flares by SPP-CNN models.

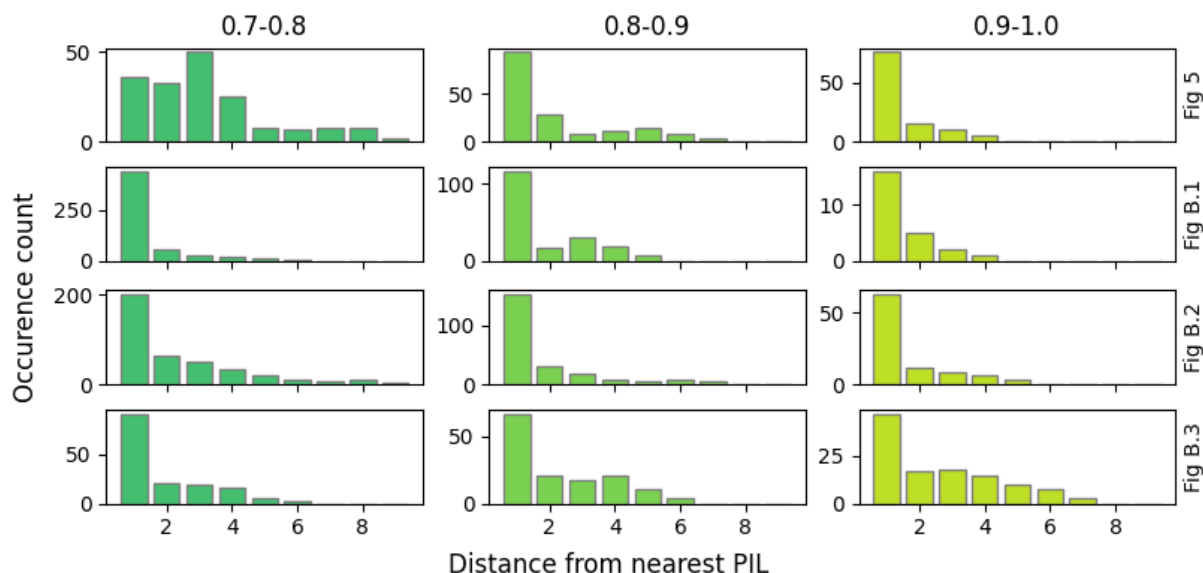


Figure 31: Histograms of the distance between pixels and their nearest PIL. The first column of plots contains histograms of the distances between pixels and their nearest PILs, using pixels in the attention range between 0.7 and 0.8 of the heatmap produced by Grad-CAM. The second column uses the attention range 0.8 to 0.9 whereas the last column uses the attention range 0.9 to 1.0.

#### 4.1.8 Limits of the current study

With this study, we can assert the strengths and weaknesses of the SPP layer through the results of our models. However, there were various drawbacks to our method, data, and labelling system. One of the major problems is the class imbalance, reflected through the low number of active regions with an X class flare. Due to the projection problem with line-of-sight magnetograms, we needed to remove every image beyond 45 degrees from the central meridian of the Sun. Unfortunately, this process removed 605 images representing 17 active regions producing X class flares. We also discarded SHARP images based on their active region also being present in other bigger files and several other data modification and selection processes during our study. After the selection and preprocessing phase, only nine distinct active regions producing X class flares remained. These pre-selection methods were not finely tuned to obtain the optimal result while keeping the most images possible. They were established on the average protocol found during our bibliographical survey. Studying which selection and preprocessing method is optimal and how they impact the training is a process we could work on in the future. Furthermore, this could help us establish a way to reliably compare studies using DL for solar flare forecasting but with different methods of handling data.

Although we obtained satisfying prediction skill scores for models using the SPP-CNN architecture, we believe further improvements can be made. A major limiting factor of this study is that we had to match as much as possible the architectures of the SPP-CNN model to that of the more traditional architectures, to compare their results and to prove the possible improvements brought by an SPP layer. This may have limited its predictive ability. Now that this comparison has been made, we can work on optimizing the SPP architecture and compare its performances with the SPP architecture of the present study.

The use of the Score-Oriented-Loss function and other loss functions can also to be studied for the specific purpose of flare prediction and to improve the prediction performance of DL models as we saw a relative improvement in the prediction ability of our Trad-CNN in predicting flares with a SOL function compared to the use of a BCE loss function.

Another improvement possible in this study is the labelling system for wider uses and better predictions. We used images with a flare in less than 24 hours as the positive class. However, there may be very little evolution in this time range, and it is reflected in the difference in accuracy for X-labelled images and FX-labelled images for the CMX classification. While the classification accuracy for X-labelled images is nearing a 100% success rate, the FX-labelled image classification rate is close to a random classification. For future studies, we could choose 48h or even 72h to label flaring images. The best solution would be to train another neural network or implement another output in our CNN, dedicated to predict the time before the eruption. Furthermore, the use of the GOES classification to label our data raised several problems. A better way to label our images or to predict flares would be based on the intensity in soft X-ray emission of each flare. In this manner, the model would not be confused anymore about which class an image belongs to. This would change the subject from a classification problem to a regression.

While line-of-sight magnetograms are easy to use for DL, they are also limited in their information. The main features our models could learn about are the shape of the polarity inversion line and the size of active regions present in the image. This limits the ability of our modes to predict flares. By using other data types (like vector magnetograms), we may be able to extract more information about

the solar flare-triggering mechanisms. Apart from changing our data source, we can also study the use of other DL methods. Convolutional neural networks are easy to use and efficient in recognizing shapes in their input, but they do not handle specific features such as the temporal evolution of their data. This feature is especially interesting for solar flare prediction since flares are mainly caused by the evolution of active regions illustrated by emerging polarities, energy buildup, and other events. Furthermore, studies using temporal evolution and presenting promising results such as Guastavino et al. (2022) are slowly emerging. In future studies, we will include this feature in our predictions by using LSTM models for solar flare prediction.

In a pure metrics-oriented view, image scaling processes seem to strongly affect the predictive ability of SPP-CNN models in a negative sense. However, we need to keep in mind that SPP-CNN was not trained using a dataset impacted by image scaling processes. The images of the training dataset of SPP-CNN were unmodified. In this sense, the model was not trained to extract compressed or distorted features, nor trained to process images of ARs that may be physically impossible.

Although the results of this scaling study are to be analysed with a critical mind, it is important to not dismiss completely the information that we can extract from it. The real variation of the predictive ability of a model trained to extract features from scaled images compared to the models used in this study may be lower than the results of this study, but we can suppose that there is still a significant loss of predictive ability. The loss of information and the distortion of physical features linked to image scaling processes impact the readability of the triggering mechanisms of solar flares. Furthermore, in the case of a model with a fixed input size, the images of the dataset used are unevenly resized and stretched. This creates a discrepancy between the size and the shape of physical features in a small image and a large image. For example, long PILs in a large image may appear of the same size as a short PIL in a small image after resizing. This invalidates any standard or scale between images and can confuse the model if two images are presented with the same physical features but are labelled differently.

## 4.2 Active region parametrization and classification with Variational Autoencoders

### 4.2.1 Active region parametrization

- **Data prepared:** We archived the Space-weather HMI Active Region Patch (SHARP) dataset, which contains images and related measured quantities from NASA's Solar Dynamics Observatory (SDO) and its Helioseismic and Magnetic Imager (HMI), so it can be used as input for further machine-learning analyses of solar active regions.
- **Models developed and validated:** We developed and thoroughly tested a class of machine-learning models ( $\beta$ -VAEs) to capture key patterns in active region magnetic-field structure and evolution, with an emphasis on producing representations that are easier to interpret.
- **Robust comparison across designs:** We built and evaluated a suite of  $\beta$ -VAE model variants on a carefully curated dataset, systematically varying key design choices—most importantly the  $\beta$  hyperparameter (which controls the strength of regularization), the number of latent dimensions (the size of the compressed representation), and the depth/complexity of the DCNN architecture (e.g., the number of hidden layers). This controlled comparison allowed us to identify configurations that are both stable across the dataset and better at separating distinct, interpretable sources of variation in active-region magnetic-field images.

- **Best-performing model applied:** We applied the strongest model to the SHARP vector magnetic-field maps (image data), compressing complex magnetic structures into a compact, interpretable set of features that supports downstream analysis and predictive modelling.

#### 4.2.2 Active region classification

- **Relationship analysis:** We examined both linear and non-linear relationships between the machine-learning-derived features and the standard SHARP parameters (see Figure 32).
- **Combined dataset:** We created an integrated dataset that merges SHARP empirical quantities describing the state of an active region's magnetic field with  $\beta$ -VAE latent-space features, enabling correlation studies and predictive modelling.
- **Curated sample:** We assembled a dataset of 50,000 SHARP images spanning 50 active regions from Solar Cycle 24 (2011–2019), together with their associated empirical metadata ( $\approx 16$  empirical SHARP parameters per time stamp).
- **Low-dimensional visualization:** We applied t-SNE and UMAP to compress the SHARP-parameter feature space ( $50,000 \times 16$ ) into a 2D representation for visualization and structure discovery (example in Figure 33). The results show that achieving a well-structured and interpretable latent space depends on a careful balance between latent regularization (set by  $\beta$ ) and the number of latent dimensions: too little regularization can yield entangled features, while too much can suppress meaningful variability, limiting how clearly different active region properties separate in the embedding.
- **Unsupervised structure discovery:** We performed unsupervised clustering and structure analysis of the  $\beta$ -VAE latent features using methods such as t-SNE (for visualization) and HDBSCAN (for density-based clustering). We carried out an equivalent analysis on the empirical SHARP parameters for comparison.
- **Evolution of latent features and active region morphology:** We used a Long Short-Term Memory (LSTM) time-series model with two LSTM cells/layers to predict the time evolution of the VAE latent dimensions, which encode key aspects of active region morphology. When we decoded these predicted latent trajectories back into images, the reconstructions matched the original observations up to  $\sim 70\%$ , indicating that the model captures a substantial fraction of the morphological evolution. This approach provides a data-driven way to track and forecast changes in active region structure, and to explore potential flare precursors reflected in evolving morphology.

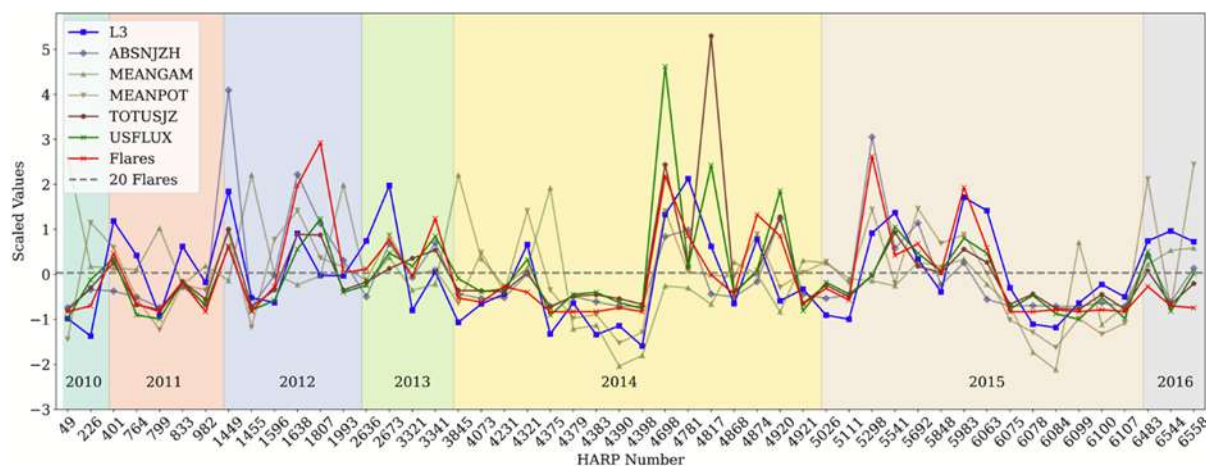


Figure 32: Time series of scaled quantities for the 50 active regions. This example highlights the relationship between the third latent variable of the same  $\beta$ -VAE model used in Figure 33 and the corresponding SHARP parameters. Particularly interesting is the closer relation between the latent variable and the number of flares.

- Feature fusion for flare prediction:** We combined geometric/morphological features learned by the  $\beta$ -VAE (from its latent space) with conventional SHARP parameters to predict flare occurrence on a held-out test dataset, linking unsupervised representation learning with supervised prediction.
- Flare now-casting method:** We used k-nearest neighbours (K-NN) for flare now-casting, trained on the combined feature set (SHARP parameters +  $\beta$ -VAE latent features) to assess how much the learned features add beyond the empirical quantities.

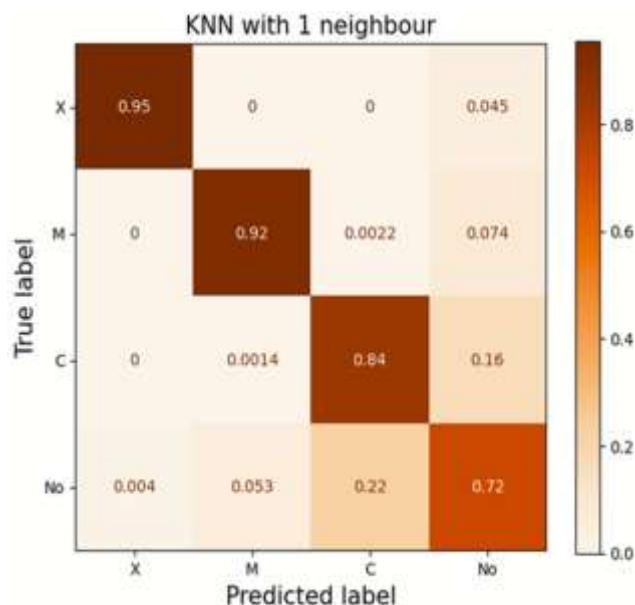


Figure 33: Example of solar-flare classification (now-casting; nomenclature follows the NOAA X-ray flare classification scheme) using a k-nearest neighbors (KNN) classifier. The results are based on the combined dataset of empirical SHARP parameters and  $\beta$ -VAE latent features, and show a good balance between precision and recall, with an F1 score of 0.88.

The current parameterization and classification scheme using VAE for feature extraction, unsupervised (t-SNE and HDBSCAN) and supervised machine learning algorithms (K-NN), shows promising results. The VAE features were successfully used to identify the SHARP images according to

the HARP number of the active region. The morphological evolution of the magnetic field in the segment can be followed using these VAE features, which can then be related to the evolution of the empirical SHARP parameters. This allows for a combined parameterization of the ARs based on ML features and man-made (empirical) values. This was achieved using only one of the available SHARP segment types, i.e., the radial component of the VMF. We recommend using the other two components, resulting in a more complete latent space representation. Furthermore, increasing the number of samples will allow for greater statistical power and better coverage of solar cycle phases.

#### **4.3 Flare prediction using CNNs on Full-Sun EUV images**

When training on SDO/AIA 193 Å data from January to August 2012, optimizing hyperparameters using validation data from September and October 2012, and testing on November and December 2012, we achieved an MSE (Mean Squared Error) of 0.34 and an  $R^2$  of 0.49, when forecasting the maximum soft X-ray flux for the next 30 minutes. Following this, we conducted CMX solar flare classification yielding an accuracy of 80% and an F1-score of 0.56. In the near future, we plan to incorporate additional data channels into the framework. We are also investigating various time windows to better map soft X-ray flux, aiming to align our work with existing literature and improve model performance by ensuring consistent training data.

#### **4.4 Related results**

This project was an opportunity for the ROB and KULeuven teams to develop their expertise and collaboration in deep learning, which continues after the end of DELPHI. And to reuse this expertise on other projects, e.g. the following publication:

- E. Tassan-Din, A. Gunessee, P. Vong, C. Marque, L. Dolla, A. Martinez, and C. Monstein: Automated Detection and Classification of Solar Radio Bursts in CALLISTO Spectrograms Using YOLOv5 and Ensemble Methods, submitted to Journal of Space Weather and Space Climate (submitted)

Although it was not the main objective of this project, the results of DELPHI can be used to develop, in the future, operational tools for flare prediction, in support of the duties of the space weather forecasters of the Royal Observatory of Belgium. Space weather is a direct societal and economical interest, illustrated by the inclusion of space weather bulletins to airplane pilots by ICAO.

#### **4.5 Recommendations**

The follow-up committee praised the quality and excellence of the work in general. They appreciated the use of Explainable AI methods to understand what features the models are using for their predictions, and found interesting that the models provide a confidence level, useful for future operational flare forecasting applications. The time dependency studies work very well. The jury was also impressed with the quality of the presentations, because it is difficult to explain Machine Learning together with solar physics and the team appears to have converged on a really effective visual strategy.

The committee also made the following remarks and recommendations.

##### **4.5.1 Flare predictions applying CNNs on line-of-sight magnetograms of individual active regions**

- To test the neural network models after cutting the magnetograms in a different way than what is done in the SHARP data set, to verify the stability of the performances.

- To add the SHARP integrated parameters in the inputs (total unsigned flux, etc...)
- To use the flaring history of the active region as input (time variations)
- To verify if for different classes of flares, the neural network look at different parts of the active region

#### **4.5.2 Solar flare prediction using CNNs on full-Sun EUV images and coronal structure segmentation**

- The site <https://lmsal.com/hek/> provides labelled AIA data (provided by SDO team) and is python-friendly. It can be useful for flare validation too.
- Filament segmentation is easier in multi-spectral data sets, particularly if the 304 Å passband is included (and which is more likely to be available for onboard implementation).

#### **4.5.3 Active region parametrization and classification with Variational Autoencoders**

- To develop an output that is useful for space weather operators
- To use the Autoencoder model to generate synthetic magnetograms of active regions that will flare and possibly compensate for the class imbalance?

Some of these recommendations have been already taken into account, or will be taken into account in the follow-up projects.

## 5. DISSEMINATION AND VALORISATION

The results of this project have been presented in internal seminars at ROB and KULeuven. They have also been presented in the following conferences:

- **Combining Physics-Derived and Machine-Learned Features for Probabilistic Solar Flare Forecasting.** K. Dineva+, ESWW 2025, Umeå, Sweden — **Oct 27–31, 2025**
- **A Multi-Perspective Comparison of Deep Learning and Traditional Machine Learning Methods for Onboard Solar Coronal Structure Detection.** P. Gonidakis+, European Space Weather Week 2025. Umea, Sweden — Oct 27–31, 2025
- **Comparing Efficiency and Performance for onboard Solar Structure Segmentation and Detection on SDO AIA Level 0 and Level 2 Data.** P. Gonidakis+, LMAG2025, Johns Hopkins University Applied Physics Laboratory, USA — 13-17 October 2025
- **Combining Physics-Derived and Machine-Learned Features for Probabilistic Solar Flare Forecasting.** K. Dineva+, 3rd ML-Helio, Madrid, Spain — **Sep 22–26, 2025**
- **Soft X-ray Flux Prediction for Onboard 24-Hour Solar Flare Forecasting Using CNNs and SDO/AIA Images.** P. Gonidakis+, MLHelio2025, Madrid, Spain — 22-26 September 2025
- **Semantic segmentation of SHARP vector magnetic field maps using Kohonen Self-Organizing Maps.** K. Dineva+, IAGA / IASPEI 2025, Lisbon, Portugal — **Aug 31–Sep 5, 2025**
- **Segmentation and characterization of solar coronal structures: Comparing the efficiency of deep learning and traditional machine learning methods.** P. Gonidakis+, IAGA / IASPEI 2025, Lisbon, Portugal — 31 August - 5 September 2025
- **Parametrization of SHARP Vector Magnetic Field Using Disentangled Representation Learning.** K. Dineva+, EGU General Assembly 2025, Vienna, Austria — **Apr 27–May 2, 2025**
- **Efficient Segmentation and Clustering of Solar Coronal Structures: A Comparison of U-Net and Classical Computer Vision Techniques Using SDO Data.** P. Gonidakis+, EGU General Assembly 2025, Vienna, Austria, EGU25-9849 — 27 April-2 May 2025
- **Bypassing the static input size of Neural Networks in flare forecasting by using Spatial Pyramid Pooling.** P. Vong+, Machine learning and Computer vision in Heliophysics (MCH) 2025, Sofia, Bulgaria, 7-9 April 2025
- **Efficient Segmentation, Detection and Clustering of Solar Coronal Structures: A Comparison of U-Net, YOLO and Classical Computer Vision Techniques Using SDO Data.** P. Gonidakis+, Machine Learning and Computer Vision in Heliophysics (MCH25), Sofia, Bulgaria — 07-09 April 2025
- **Flare Forecasting Framework Based on SHARP Features Extracted with Disentangled Representation Learning.** K. Dineva+, 20th European Space-Weather Week (ESWW2024), Coimbra, Portugal — **Nov 4–8, 2024**

- **Segmentation and Feature Engineering of Solar Coronal Structures Using SDO AIA and HMI Data.** P. Gonidakis+, European Space Weather Week 2024, Coimbra, Portugal — 4-8 November 2024
- **Bypassing the static input size of Neural Networks in flare forecasting by using Spatial Pyramid Pooling.** P. Vong+, ESWW 2024, Coimbra, Portugal — 4-8 November 2024
- **Parametrization of SHARP Vector Magnetic Field Using Disentangled Representation Learning.** K. Dineva+, 17th European Solar Physics Meeting (ESPM-17), Turin, Italy — **Sep 9–13, 2024**
- **Predicting Soft X-ray Emissions for Solar Flare Forecasting Using a Self-Supervised CNN Trained on Solar Dynamics Observatory Data.** P. Gonidakis+, 17th European Solar Physics Meeting ESPM-17, Turin, Italy — 9-13 September 2024
- **Machine Learning Based Parametrization of Solar Active Regions Using Disentangled Variational Autoencoders.** K. Dineva+, Machine Learning for Astrophysics 2nd Edition (ML4ASTRO2), Catania, Italy — **Jul 8–12, 2024**
- **Investigation of the VAE and Their Potential for Active Region Classification and Flare Prediction.** K. Dineva+, 19th European Space-Weather Week (ESWW2023), Toulouse, France — **Nov 20–24, 2023**

The data and methodology of this project have been the basis of topics of Master thesis for KULeuven students:

1. [\*Classification of Solar Energetic Activity using Data Analysis and Clustering of Active Regions\*](#), Hanne Baeke, KULeuven, 2021
2. [\*Autoencoding of Solar Images and Generation of Artificial Active Regions\*](#), Orestis Angelos Karapiperis, KULeuven, 2021

The results of this project had and will have a direct impact on the strategic scientific objectives of the Royal Observatory of Belgium and KULeuven by producing a new set of methods and knowledge that are essential for the operation of our Space Weather services. In particular:

- the methodology used for the flare prediction with CNNs has been also applied for classification of radio bursts at ROB (Tassan-Din et al., submitted, see publications above);
- the segmentation methodology will be used to feed other neural networks for flare prediction (e.g. to extract cutout images);
- The same methodology can be used to develop onboard autonomy in spacecraft operated for space weather monitoring, for example, after detection of an eruption, the autonomous triggering of special measurement modes and the selective downlink of plasma environment parameters; the advanced on-board data analysis of three-dimensional particle distribution

functions; the on-board analysis of solar images; the on-board prediction capability of Solar Energetic Particle related hazards. This aspect has been presented in the following conference presentation:

**Machine Learning Algorithms for Autonomous Space Mission Operations.** Tommaso Torda, Tommaso Alberti, Giuseppe Consolini, Rossana De Marco, Ekaterina Dineva, Jonah Ekelund, Panagiotis Gonidakis, Monica Laurenza, Maria Federica Marcucci, Stefano Markidis, George Miloshevich, Stefaan Poedts, Begnamino Sanò, Nicolina Chrysaphi, EGU General Assembly 2025, Vienna, Austria, EGU25-16713 — 27 April-2 May 2025

Moreover, this project has been the seed of the Ph.D. project of Philippe Vong (started in September 2024) to build on the developed methodology and improve our flare prediction capabilities.

Finally, our ultimate goal is to develop operational flare forecasting tools.

This project will lead to major improvements in the knowledge of the sources of space weather having an impact on our planet, and pave the way for the development of new space weather operational tools. We are developing additional modern capacities and skills that will enhance our international competitiveness. The academic and operational advances will also have a practical impact on the economy and the civil society: space weather directly affects GPS positioning, satellite communications, high altitude human radiation, and electricity distribution among others. An improvement on our ability to forecast high energy transfers from the Sun is crucial to these industries.

## 6. PUBLICATIONS

Non peer-reviewed articles:

1. P. Antunes, M.I. Al Hafiz, J. Ekelund, E. Dineva, G. Miloshevich, P. Gonidakis & A. Podobas (2025): *Evaluating four fpga-accelerated space use cases based on neural network algorithms for on-board inference*, In Proceedings of the 18th IEEE International Symposium on Embedded Multicore/Manycore SoCs (MCSOC), Singapore.

Peer-reviewed articles:

1. P. Vong, L.R. Dolla, A. Koukras, J. Gustin, J. Amaya, E. Dineva, G. Lapenta (2025): *Bypassing the static input size of neural networks in flare forecasting by using spatial pyramid pooling*, A&A, 695, A65, [10.1051/0004-6361/202449671](https://doi.org/10.1051/0004-6361/202449671)
2. H. Baeke, G. Miloshevich, E. Dineva, F. Carella, P. Gonidakis, J. Amaya and G. Lapenta (2025): *Classifying Solar Flares with kNN on Sparse Autoencoder Representations*, Monthly Notices of the Royal Astronomical Society, <https://doi.org/10.1093/mnras/staf2271>
3. P. Gonidakis, F. Carella, E. Dineva, H-J. Jeong, P. Antunes, A. Podobas, S. Raptis, V. Toy-Edens, M. Jin, S. Poedts, J. Magdalenic, G. Miloshevich: *Comparing Solar Structure Detection Methods in SDO/AIA observations and the application to raw uncalibrated data*, submitted to JGR: Machine Learning and Computation
4. E. Dineva, I. Kontogiannis, G. Lapenta, G. Miloshevich, J. Amaya, J. Magdalenic, and S. Poedts: *Parametrization of Active Region Vector Magnetic Field Using Disentangled Representation Learning*, to be submitted to A&A.

## 7. ACKNOWLEDGEMENTS

We thank the members of the followup committee for their support and useful suggestions:

1. Pr. Benoit Frénay (University of Namur, Belgium)
2. Jesse Andries (ROB; World Meteorological Organization)
3. Barbara J. Thompson (Solar Physics Laboratory/NASA, Greenbelt, USA)
4. Simon Wing (The John Hopkins University, Applied Physics Laboratory; Andrews University, Physics Department).

We also thank Jorge Amaya and Alexandros Koukras who substantially contributed to the project proposal before flying to new horizons, while still providing discussion and support, and George Miloshevich for useful discussion and critical viewing on the results of our studies.

This work is dedicated to Giovanni Lapenta from KULeuven, who co-promoted this project before his sudden death. We miss you, Gianni!

## BIBLIOGRAPHIC REFERENCES

- Agarap, Abien Fred (2018): *Deep Learning using Rectified Linear Units (ReLU)*, ArXiv, <https://doi.org/10.48550/arXiv.1803.08375>
- Ahmed O.W., Rami Qahwaji, Tufan Colak, Paul A. Higgins, Peter T. Gallagher & D. Shaun Bloomfield (2013): *Solar Flare Prediction Using Advanced Feature Extraction, Machine Learning, and Feature Selection*, Sol Phys 283: 157. <https://doi.org/10.1007/s11207-011-9896-1>
- Benz A.O. (2017): *Flare Observations*, Living Rev. Sol. Phys. 14: 2. <https://doi.org/10.1007/s41116-016-0004-3>
- Bloomfield D. Shaun, Paul A. Higgins, R. T. James McAteer, and Peter T. Gallagher (2012): *Toward Reliable Benchmarking of Solar Flare Forecasting Methods*, The Astrophysical Journal Letters, Volume 747, Issue 2, article id. L41
- Bobra M. G., X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes & K. D. Leka (2014): *The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-Weather HMI Active Region Patches*, Sol. Physics 289, 3549. <https://doi.org/10.1007/s11207-014-0529-3>
- Bobra M. G., and S. Couvidat (2015): *Solar Flare Prediction Using SDO/HMI Vector Magnetic Field Data with a Machine-learning Algorithm*, ApJ 798, 135
- Camporeale, E. (2019): *The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting*, Space Weather 17, Issue 8, pp. 1166. <https://doi.org/10.1029/2018SW002061>
- Devos A., Verbeeck C. and Robbrecht E. (2014): *Verification of space weather forecasting at the Regional Warning Center in Belgium*, Space Weather Space Clim. 4, A29. <https://doi.org/10.1051/swsc/2014025>
- Dudík J., Vanessa Polito, Miho Janvier, Sargam M. Mulay, Marian Karlický, Guillaume Aulanier, Giulio Del Zanna, Elena Dzifčáková, Helen E. Mason, and Brigitte Schmieder (2016): *Slipping Magnetic Reconnection, Chromospheric Evaporation, Implosion, and Precursors in the 2014 September 10 X1.6-Class Solar Flare*, The Astrophysical Journal 823(1) <https://doi.org/10.3847/0004-637X/823/1/41>
- Falconer A.D., Moore, Ronald L., Barghouty, Abdunnasser F. , Khazanov, Igor (2012): *Prior Flaring as a Complement to Free Magnetic Energy for Forecasting Solar Eruptions*. ApJ 757, 32. <https://doi.org/10.1088/0004-637X/757/1/32>
- Fletcher L., Dennis, B. R., Hudson, H. S. (2011): *An Observational Overview of Solar Flares*. Space Sci Rev 159: 19. <https://doi.org/10.1007/s11214-010-9701-8>
- Florios K., Kontogiannis, Ioannis, Park Sung-Hong, Guerra Jordan A., Benvenuto, Federico; Bloomfield, D. Shaun; Georgoulis, Manolis K. (2018): *Forecasting Solar Flares Using Magnetogram-based Predictors and Machine Learning*. Sol Phys 293: 28. <https://doi.org/10.1007/s11207-018-1250-4>
- Galvez Richard, Fouhey David F., Jin Meng, Szenicer Alexandre, Muñoz-Jaramillo Andrés, Cheung Mark C. M., Wright Paul J., Bobra Monica G., Liu Yang, Mason James, and Thomas Rajat (2019): *A Machine-learning Data Set Prepared from the NASA Solar Dynamics Observatory Mission*, The Astrophysical Journal Supplement Series, Volume 242, Issue 1, article id. 7
- Guastavino, Sabrina, Marchetti, Francesco, Benvenuto, Federico, Campi, Cristina, Piana, Michele: *Implementation paradigm for supervised flare forecasting studies: A deep learning application with video data*, Astronomy & Astrophysics, Volume 662, id.A105
- He K., Zhang, X., Ren, S., Sun, J. (2014): *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8691. Springer, Cham. [https://doi.org/10.1007/978-3-319-10578-9\\_23](https://doi.org/10.1007/978-3-319-10578-9_23)

- Huang, Xin, Wang, Huaning, Xu, Long, Liu, Jinfu, Li, Rong, Dai, Xinghua (2018): *Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms*, The Astrophysical Journal, Volume 856, Issue 1, article id. 7. <https://doi.org/10.3847/1538-4357/aaae00>
- Ioffe Sergey and Christian Szegedy (2015): *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, ArXiv, <https://doi.org/10.48550/arXiv.1502.03167>
- Jing Ju, Song Hui, Abramenko Valentyna, Tan Changyi, Wang, Haimin (2006): *The Statistical Relationship between the Photospheric Magnetic Parameters and the Flare Productivity of Active Regions*, The Astrophysical Journal, Volume 644, Issue 2, 1273
- Jolliffe, I. T. (2002): *Principal Component Analysis*. 2nd ed. Springer Series in Statistics. New York, NY: Springer. <https://doi.org/10.1007/b98835>
- Jonas E., Monica Bobra, Vaishaal Shankar, J. Todd Hoeksema & Benjamin Recht (2018): *Flare Prediction Using Photospheric and Coronal Image Data*. Sol Phys 293: 48. <https://doi.org/10.1007/s11207-018-1258-9>
- Kingma, Diederik P., and Max Welling (2019): *An Introduction to Variational Autoencoders*, Foundations and Trends in Machine Learning 12, no. 4 : 307–92. <https://doi.org/10.1561/22000000056>
- Korsós M. B., Baranyi, T. ; Ludmány, A. (2014): *Pre-flare Dynamics of Sunspot Groups*. ApJ 789, 107. <https://doi.org/10.1088/0004-637X/789/2/107>
- Korsós B. M., Chatterjee, P.; Erdélyi, R. (2018): *Applying the Weighted Horizontal Magnetic Gradient Method to a Simulated Flaring Active Region*. ApJ 857, 103. <https://doi.org/10.3847/1538-4357/aab891>
- Korsós M.B.; Gyenge, N. ; Baranyi, T. ; Ludmány, A. (2015): *Dynamic Precursors of Flares in Active Region NOAA 10486*. J Astrophys Astron 36: 111. <https://doi.org/10.1007/s12036-015-9329-x>
- LeCun, Y.A., Bottou, L., Orr, G.B., Müller, KR. (2012): *Efficient BackProp*. In: Montavon, G., Orr, G.B., Müller, KR. (eds) *Neural Networks: Tricks of the Trade*. Lecture Notes in Computer Science, vol 7700. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-35289-8\\_3](https://doi.org/10.1007/978-3-642-35289-8_3)
- Leka K. D. and Barnes G. (2003): *Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. II. Discriminant Analysis*. ApJ 595, 2, pp. 1296-1306. doi:10.1086/377512
- Leka K. D. et al. (2019): *A Comparison of Flare Forecasting Methods. II. Benchmarks, Metrics, and Performance Results for Operational Solar Flare Forecasting Systems*. The Astrophysical Journal Supplement Series 243, Issue 2, 36. doi:10.3847/1538-4365/ab2e12
- Leka K. D. et al. (2019b): *A Comparison of Flare Forecasting Methods. III. Systematic Behaviors of Operational Solar Flare Forecasting Systems*. The Astrophysical Journal 881, Issue 2, 101. doi:10.3847/1538-4357/ab2e11
- Li, Xuebao, Zheng, Yanfang, Wang, Xinshuo, Wang, Lulu (2020): *Predicting Solar Flares Using a Novel Deep Convolutional Neural Network*, The Astrophysical Journal, Volume 891, Issue 1, id.10
- Liu Hao, Chang Liu, Jason T. L. Wang, and Haimin Wang (2019): *Predicting Solar Flares Using a Long Short-term Memory Network*. ApJ 877, 121. <https://doi.org/10.3847/1538-4357/ab1b3c>
- Mackovjak, Šimon, Martin Harman, Viera Maslej-Krešňáková, Peter Butka (2021): *SCSS-Net: solar corona structures segmentation by deep learning*, Monthly Notices of the Royal Astronomical Society, Volume 508, Issue 3, Pages 3111–3124, <https://doi.org/10.1093/mnras/stab2536>
- Marchetti F., Guastavino S., M. Piana, C. Campi (2022): *Score-Oriented Loss (SOL) functions*, Pattern Recognition, Volume 132, article id. 108913.

- McInnes, Leland, John Healy, and James Melville (2018), *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv preprint arXiv:1802.03426. <https://doi.org/10.48550/arXiv.1802.03426>.
- McIntosh Patrick S. (1990): *The classification of sunspot groups*. Solar Physics, vol. 125, p. 251-267. doi:10.1007/BF00158405
- Nishizuka N., K. Sugiura, Y. Kubo, M. Den, S. Watari, and M. Ishii (2017): *Solar Flare Prediction Model with Three Machine-learning Algorithms using Ultraviolet Brightening and Vector Magnetograms*. ApJ. 835, 156. <https://doi.org/10.3847/1538-4357/835/2/156>
- Poisson Mariano, Démoulin, Pascal, López Fuentes, Marcelo, Mandrini, Cristina H. (2016): *Properties of Magnetic Tongues over a Solar Cycle*, Solar Physics, Volume 291, Issue 6, pp.1625
- Schrijver J. Carolus (2007): *A Characteristic Magnetic Field Pattern Associated with All Major Solar Flares and Its Use in Flare Forecasting*. The Astrophysical Journal, Volume 655, Issue 2, pp. L117-L120. doi:10.1086/511857
- Schwenn R. (2006): *Space Weather: The Solar Perspective*. Living Rev. Sol. Phys. 3: 2. <https://doi.org/10.12942/lrsp-2006-2>
- Selvaraju, Ramprasaath R., Cogswell, Michael, Das, Abhishek, Vedantam, Ramakrishna, Parikh, Devi, Batra, Dhruv (2016): *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*, [arXiv:1610.02391](https://arxiv.org/abs/1610.02391)
- Shibata K. and Magara T. (2011): *Solar Flares: Magnetohydrodynamic Processes*. Living Rev. Sol. Phys. 8: 6. <https://doi.org/10.12942/lrsp-2011-6>
- Srivastava Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov (2014): *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research 15(56):1929
- SunPy Community, Barnes, Will T., Bobra, Monica G., et al. (2020): *The SunPy Project: Open Source Development and Status of the Version 1.0 Core Package*, The Astrophysical Journal, Volume 890, Issue 1, id.68
- van der Maaten, Laurens (2014): *Accelerating t-SNE using Tree-Based Algorithms*, Journal of Machine Learning Research 15, no. 93: 3221–3245. <https://www.jmlr.org/papers/volume15/vandermaaten14a/vandermaaten14a.pdf>
- Wang Haimin, Jing Ju, Tan, Changyi, Wiegelmann, Thomas, Kubo, Masahito (2008): *Study of Magnetic Channel Structure in Active Region 10930*, The Astrophysical Journal, Volume 687, Issue 1, pp. 658-667
- Wang Xiantong, Yang Chen, Gabor Toth, Ward B. Manchester, Tamas I. Gombosi, Alfred O. Hero, Zhenbang Jiao, Hu Sun, Meng Jin, and Yang Liu (2020): *Predicting Solar Flares with Machine Learning: Investigating Solar Cycle Dependence*. ApJ 895, 3. <https://doi.org/10.3847/1538-4357/ab89ac>
- Welsch B. T., Yan Li, Peter W. Schuck, and George H. Fisher (2009): *What is the relationship between photospheric flow fields and solar flares?* ApJ 705, 821. <https://doi.org/10.1088/0004-637X/705/1/821>
- Wheatland M.S. (2004): *A Bayesian Approach to Solar Flare Prediction*. ApJ 609, 1134. <https://doi.org/10.1086/421261>
- Yamashita, R., Mizuho Nishio, Richard Kinh Gian Do and Kaori Togashi (2018): *Convolutional neural networks: an overview and application in radiology*. Insights Imaging 9, 611–629 . <https://doi.org/10.1007/s13244-018-0639-9>

Yuan Yuan, Shih, Frank Y. ; Jing, Ju ; Wang, Hai-Min (2010): *Automated flare forecasting using a statistical learning technique*. Res. Astron. Astrophys. Volume 10, page 785.

<https://doi.org/10.1088/1674-4527/10/8/008>

Zhou, Zongwei, Vatsal Sodha, Jiaxuan Pang, Michael B. Gotway, Jianming Liang (2021): *Models genesis*, Medical image analysis 67, 101840.

Zirin H. and Wang H. (1993): *Narrow lanes of transverse magnetic field in sunspots*. Nature, volume 363, pages 426–428. <https://doi.org/10.1038/363426a0>