Limits of solar flare forecasting models and new deep learning approach

G. Francisco[1,2,3] M. Berretti[1,4] S. Chierichini[5,1,3] R. Mugatwala[1,5,3] J. Fernandes[6] T. Barata[2] D. Del Moro[1]

[1]Department of Physics, University of Rome Tor Vergata, Rome, Italy
[2]IA, Instituto De Astrofisica E Ciências Do Espaço, University of Coimbra, Coimbra, Portugal
[3]Department of Physics, University of Rome La Sapienza, Rome, Italy
[4]University of Trento, Trento, Italy
[5]SP2RC, School of Mathematics and Statistics, University of Sheffield, Sheffield, UK
[6]CITEUC, Geophysical and Astronomical Observatory, University of Coimbra and Department of Mathematics, Coimbra, Portugal

Key Points:

- Flare forecasting models reaching good performances may still struggle to detect activity change and outperform Persistence models
- New metrics are defined to highlight model weaknesses better and ease comparisons
- Patch-Distributed-CNNs provide state-of-the-art full-disk forecasts with position information and sub-regional risk assessment

Corresponding author: G. Francisco, gregoire.francisco@gmail.com

Abstract

Reliable forecasting models are necessary to mitigate the risks posed by solar flares to human technology. This study introduces a novel deep learning forecasting approach while emphasizing the need for performance evaluation methods tailored to better highlight current models' limitations. We discuss shortcomings in existing evaluation metrics such as the True Skill Statistic (TSS) and the Heidke Skill Score (HSS) and propose the Matthews Correlation Coefficient (MCC) as a consistent alternative to both of them. We introduce Persistence-Relative-Skill Scores (PRSS), as well as metrics evaluation on the restriction of time windows presenting a change of activity (Activity-Changes (AC) performances). Models reaching state-of-the-art performances with traditional metrics can struggle to be more efficient and explanatory than a simple Persistence model and to forecast change in activity significantly better than random guesses. We introduce the Patch-Distributed-CNNs (P-CNN), which allow to perform full-disk forecasts while providing event probabilities in solar sub-regions and position predictions. This new framework offers similar information to Active-Region-based forecasting models while bypassing the problem of unrecorded and misattributed flares that are detrimental to machine learning training. As a result, the model also operates independently of prior feature extraction and AR detection, thus offering promising operational utility with minimal external dependencies. Finally, a method is proposed for constructing balanced and independent Cross-Validation folds for full-disk models. Models combining SDO/AIA EUV images as inputs show improved performances compared to employing SDO/HMI photospheric magnetograms, with a TSS of 0.74 for the C+ model and 0.62 for the M+ model.

Plain Language Summary

Flare forecasting models deemed as state-of-the-art with standard performance evaluation methods can exhibit poor skills in forecasting changes in activity. They barely compete with Persistence models, both from a practical operational point of view (F1-score) and in their ability to explain the observed events (MCC). We propose Persistence-Relative metrics and performance evaluation on the subset of time windows presenting a change of activity to highlight these models' flaws better and ease comparisons. We also propose new deep-learning models that do not rely on previously identified Active Regions while still providing position estimations for the forecasted events and predictions at regional levels of the solar disk. This new model offers promising operational utility thanks to minimal external dependencies, and more reliable training by minimizing mislabels during the training process.

1 Introduction and Related Work

1.1 Deep learning To forecast flares

Solar flares are one of the most energetic manifestations of the solar activity. They are bursts of electromagnetic radiations and particles, believed to be caused by magnetic reconnections converting huge amounts of magnetic energy into heat and kinetic energy. To characterize the potential danger they represent, flares are commonly classified into 5 classes according to their Soft X-Rays (SXR) (wavelength from 0.1 to 0.8 nanometers) Maximum Peak Flux (MPF): A-Flares with a MPF $< 10^{-7}\ W \cdot m^{-2}$, B-Flares for a MPF $\in [10^{-7}, 10^{-6}[\ W \cdot m^{-2}$, C-Flares for a MPF $\in [10^{-6}, 10^{-5}[\ W \cdot m^{-2}$, M-Flares for a MPF $\in [10^{-5}, 10^{-4}[\ W \cdot m^{-2}$ and finally X-Flares with a MPF $> 10^{-4}\ W \cdot m^{-2}$. Flares above the M-class start representing a threat to human health and technologies. This motivates a significant work effort to predict them successfully. Recently, many approaches have been proposed to forecast solar flares using deep learning. Huang et al. (2018), Park et al. (2018), Li et al. (2020), Z. Deng et al. (2021) and Pandey et al. (2023) used Convolutional Neural Network (CNN) on magnetograms images. Nishizuka et al. (2018) used an Multi-Layer Perceptron (MLP) on a combination of physical features. Guastavino et al. (2022a) combined CNN with Long Short-Term Memory (LSTM) to process temporal-stacks (videos) of magnetograms, while Deshmukh et al. (2022) combined CNN on magnetograms images with a Random Forest algorithm on some physical features to decrease the False Alarm Ratio (FAR) of their model. A study by Sun et al. (2022) used both CNN on magnetograms and LSTM on a time series of physical features and ensembled the two approaches for optimal results. The typical goal of these methods is to forecast the likelihood of the whole Sun or of a particular Active Region (AR) to produce a flare in a specific time window - often 24h - which starts from the data used for the prediction, or eventually a few minutes after. The prediction is usually binary, indicating whether there will be at least one flare above a given threshold during the forecasted window. This threshold is often the one of C or M-class flares, while the X-class is addressed less frequently due to the challenges met by machine learning methods on extremely imbalanced and scarce data. An advantage of most deep learning approaches compared to other Machine Learning (ML) works is that they do not rely on previously identified features. This is often also seen as a drawback referred to as the black-box problem (Camporeale (2019)) as we have less understanding about the way those models make a given prediction. This can nevertheless be mitigated through explainability methods. Yi et al. (2021) used such methods to show that their deep learning model was learning to extract features highly correlated to known predictive features. However, such models are not limited to known flare predictive features and could help identify new ones. As these models do not depend on pre-identified features, they also enable the development of more autonomous systems, potentially simplifying deployment and maintenance in operational settings. However, most current works focus on forecasting flare directly on pre-identified AR. Such models still suffer from external dependency and could not forecast a flare in the case of a mis-detected or un-emerged AR. van der Sande et al. (2022) also showed that many flares are unrecorded or misattributed at the AR level in standard flare catalogs. They found up to 8% of misattributed labels at the AR level and about 20% of missing events among those located between ± 65 heliographic longitude from disk center and above the M-class thresholds between 2010 and 2017, from the standard Goes Flare catalog. While van der Sande et al. (2022) acknowledge that the Heliophysics Event Knowledgebase (HEK) correct many of the missing events from the Goes Flare catalogue, several are still attributed to a wrong AR or simply lack an attribution. This leads to a significant number of ARs inaccurately labelled as events and no-events, which can impair model training and evaluation in an already challenging context of imbalanced and scarce data. Park et al. (2018), Yi et al. (2021), and Pandey et al. (2023) train their model to forecast flares directly at the whole solar disk level. Such models do not suffer from mislabels due to missing or misattributed AR. They are also more autonomous than AR-based models. However, they do not provide information about the position of the predicted flares. In this work, we propose the first deep learning model performing flare forecast at the whole disk

level from which it is possible to retrieve the position of the foreseen flares along with sub-regional predictions without relying on prior identifications of AR. We focus on forecasts within 24-hour time windows and use the whole-disk exhaustive flare catalogue from Plutino et al. (2023). Throughout this work, models performing binary forecasts above the C-class and M-class models will be designated as C+ models and M+ models, respectively.

1.2 Models evaluation

Although most of the previous literature focuses on the same problem, it is hard to objectively compare the performances of these different works without a proper benchmark dataset. Barnes et al. (2016) and Leka et al. (2019) showed that the performances of forecasting methods evaluated during the traditional training and test phases of existing flare forecasting studies are often over-optimistic and reflect poorly the performance a model would have in an operational setting. These gaps between expected and actual performances happen because the data on which a model is evaluated is typically biased and does not accurately represent the data that can be met in operational context (Cinto et al. (2020)). For instance, Guastavino et al. (2022a) identified subcases of negatives (quiet time-windows) on which models perform differently, highlighting that the performances' evaluations are strongly sensitive to their relative representation in the dataset. Guastavino et al. (2022) also introduced the distinction between good and bad errors in flare forecasting to better account for the problem's dynamical nature. Guastavino et al. (2022b) showed that using metrics weighted to distinguish such errors helps in selecting models that reach better operational performance. This work proposes two new methods to assess flare forecasting models' performances. These new evaluation methods enable highlighting that models may struggle to outperform a simple Persistence model, with remarkably low skills in forecasting activity changes, while they might still appear as efficient and useful models with conventional evaluation metrics. They also appear to partially mitigate the impact of the evaluation set compositions on the measured performance, thus helping when comparing different models on different datasets.

2 Metrics

2.1 Standard Evaluation Methods

2.1.1 Common metrics

This section introduces the most common evaluation metrics used in flare forecasting and discusses some of their shortcomings and advantages. Formulas can be found in Appendix A1, and additional insights on their discussed properties are detailed in Appendix A2

The selection and evaluation of models heavily rely on the metrics employed to gauge their performance. However, the scientific community lacks consensus regarding a singular, comprehensive metric for binary classification. Different metrics serve distinct objectives and address diverse problems. While multiple metrics are often necessary to provide comprehensive insights into a model's qualities, a single metric remain essential for model selection during training and for facilitating models comparison. In Space Weather, such as in the case of flare forecasts, evaluating models is particularly challenging due to the prevalent class imbalance between positive and negative events in operational settings.

Binary metrics are typically defined as a function of the confusion matrix (CM) (Equation A1), also referred to as the prediction-observation contingency table. From this matrix can be derived four basic rates which together summarize the information contained by the confusion matrix and a model's qualities:

- Class-Accuracy rates:
  - True Positive Rate (TPR), also known as Recall (Equation A2)
  - True Negative Rate (TNR), also known as Sensitivity (Equation A3)
- Class-Precision rates:
  - Positive Predictive Value (PPV), also known as Precision (Equation A5)
  - Negative Predictive Value (NPV)(Equation A4)

In Weather and Space Weather forecasting, models predicting extreme events typically serves to trigger weather alerts warning users in advance of potentially dangerous events. Such systems designed to trigger alerts, could be refereed as alarm systems. For such systems, the positive class corresponds to an event and the negative one to a no-event. When designing and evaluating an alarm system, a particular emphasis can typically lie on achieving high recall (TPR) to detect a maximum of events, along with high precision (PPV) to ensure confidence in positive predictions. Equivalently to the precision, practitioners alternatively look at the False Alarm Ratio (FAR) which is the complementary of the former and gives the rate at which positive predictions give a false alarm.

The F1-score (Equation A6), is the harmonic mean of the recall and the precision. It offers a consolidated assessment of an alarm system skill. It can be extended to the Fβ-score to give β-times more importance to the recall relatively to the precision, to consider defined preferences between positive events detection rate and False Alarm Ratio.

The True Skill Statistic (TSS) (Equation A7) is among the most used metrics to evaluate flare forecasting models but suffers from important limitations. In ideal cases, the TSS is independent of the class balance. However, for flare forecasting, the TSS is among the most sensitive metrics to this bias as well as other ones, both theoretically (Appendix A21) and empirically (Appendix B). The TSS synthesizes the class-accuracy rates (TPR and TNR) and, therefore, contains no information about a model's precision. This makes it arguably unsuitable to evaluate an alarm system in the case of imbalanced datasets. A very high TSS may indeed hide impractical models in imbalanced cases, as described in appendix A22. Finally, it can be noted that the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC), another common metric, can be expressed in terms of the expected value of the TSS for different probability decision thresholds. The AUC ROC thereby suffers from the same limitations as the TSS, mainly an absence of information about class-precisions, which can be detrimental in imbalanced problems.

The Heidke Skill Score (HSS) (Equation A10) is another common metric in flare forecasting and is typically used to compare a model to random guessing. It synthesizes (c.f Appendix A17) both the class-accuracy rates (TPR and TNR) and the class-precision rates (PPV and NPV) making it more complete than the TSS. However, the weight given to the class accuracies and the class precisions in the HSS is model-dependant as it relies on the model's Negative-Frequency-Bias (NFB) (Appendix A17 & Appendix A23). This makes the HSS difficult to interpret and potentially unreliable when comparing different models.

The Matthews Correlation Coefficient (MCC) (Equation A11), is currently less common in Space Weather studies but presents several advantages over the TSS and HSS (c.f. Appendix A24). It is the binary Pearson correlation coefficient between the predictions and the observations. As such the MCC directly measures how much of the observed variance of the target can be explained by the model. The MCC is, therefore, an agnostic measure of a model's explanatory power. It also synthesizes the class accuracy and precision rates by giving even weights to each of the four basic confusion matrix rates (TPR, TNR, PPV, NPV). It allows us to compare a model to random guessing in a simpler way than the HSS and has been several times shown to be one of the most complete and reliable metrics to assess a model's overall quality in classification. From our empirical observations, the MCC also appeared more resilient to dataset composition changes than the TSS or the HSS (c.f. Appendix B).

Ultimately, no single metric is enough to give a complete insight into a model's flaws and qualities. Still, the F1-Score, or alternatively an Fβ-score, remains of strong interest for selecting models intended as alarm systems. Meanwhile, the MCC appears as the most informative and reliable metric to assess the general ability of a model to explain a given problem.

2.1.2 Thresholding

In binary classification, a prevalent approach involves determining the probability threshold that maximizes a desired metric. Yet, studies have shown that optimal thresholds are usually dependent on the class-balance of the datasets (Leka et al. (2019)). This introduces an additional bias in the metric evaluation. As a result, in our evaluations, we assess all metrics using a consistent probability threshold of 0.5.

2.2 Identifying flare forecasting models weakness

The previous metrics do not allow to identify key weaknesses of current flare forecasting models. Therefore, we introduce complementary evaluation methods to better highlight important models' limitations.

2.2.1 Activity-Change (AC) and No-Change (NC) performances

We define changing time windows as the ones exhibiting an *Activity-Change (AC)*, i.e. when the window label differs from the previous consecutive non-overlapping window. Conversely, a constant time window is a window presenting *No-Change (NC)* of activity with respect to the previous one. Figure 1 display examples of the of AC and NC windows in the case of M+ forecasts.

This distinction is crucial as forecasting flares on constant versus changing time windows represents two distinct challenges. While the former is more of a classification/recognition problem, the real challenge of flare forecasting lies in the latter where current models struggle to perform significantly beyond random classifiers.

This lack of performance on changing time-window is not reflected by usual metrics like the TSS and HSS that can still reach good values on the whole evaluation set. To understand this limitation better and what type of model or information could allow to push that boundary, we propose to complement traditional performance evaluation by evaluating models' performances separately on changing and constant time-windows.

The restriction of a metric evaluated on changing time windows will be referred to here as Activity-Change-metric (AC-metric), and the restriction of a metric evaluated on constant time windows will be referred to as No-Change-metric (NC-metric).

2.2.2 Persistence Relative Skill Scores

Usual skill scores compare model performances against constant or random guess classifiers. In the context of flare forecasting, a Persistence model — predicting labels based on the previous time window — can also be seen as a no-skill model. We find in this work that such a model typically outperforms the other no-skill models and perform similarly to our best models (Section 4.1 & Appendix B2). Skill scores relative to the Persistence model could thus better highlight such limitations and the practical utility of flare forecasting models. We also find that skill scores defined relatively to the Persistence model can be less sensitive to the dataset composition (c.f. Appendix B2), which could make them more suited to compare flare forecasting models.

Example of AC and NC windows for 24H M+ binary forecasts

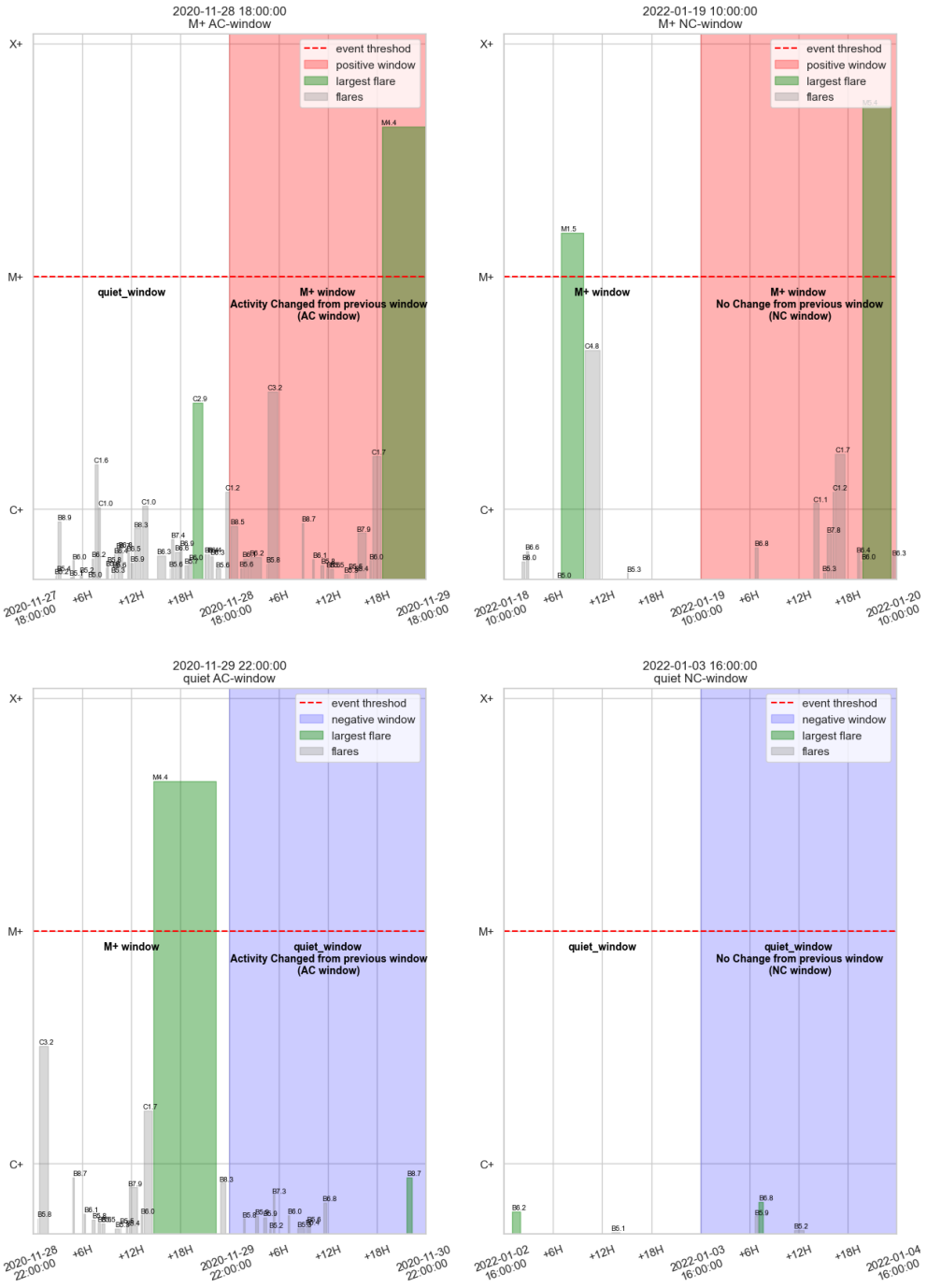


Figure 1: First column displays examples of AC-windows, with a positive AC-window on the first row (colored in red), and a negative AC-window on the second row (colored in blue). Second column displays examples of NC-windows: positive NC-window on the first row and negative NC-window for the second one. The white time-windows are the one that precede the windows of interest. The gray bars correspond to flares plotted from their starting to end dates. The green bars correspond to the biggest flare inside the corresponding window. Label on top of the flares correspond to their SXR-MPF.

For a metric S defined on $[Inf, Sup] \subset \mathbb{R}$, we define the Persistent Relative Skill Score (PRSS) as the difference of a model's score (S_{model}) with a Persistence model (S^*), rescaled in $[-1, 1]$, according to the following equation :

$$PRSS = \begin{cases} \frac{S_{model} - S^*}{Sup - S^*} & \text{if } S_{model} \geq S^* \\ \frac{S_{model} - S^*}{S^* - Inf} & \text{if } S_{model} < S^* \end{cases} \quad (1)$$

It can be noted that in both cases, respectively, over- and underperforming, the denominator represents the maximum possible overperformance or underperformance with respect to the Persistence score. Consequently, when positive, the PRSS reads as the percentage of the maximum possible overperformance - respectively underperformance when negative - achieved by the evaluated model relatively to the Persistence. Finally, the PRSS is null if and only if $S_{model} = S^*$, it equals 1 when $S_{model} = Sup$, and equals -1 when $S_{model} = Inf$.

It is noted that these metrics diverge from common binary classification metrics, as they factor in parameters external to the model confusion matrix by including terms of the Persistence model's confusion matrix.

3 Method

3.1 Data

A dataset was prepared for model training and evaluation, encompassing 56,664 samples from May 14, 2010, to April 18, 2023, with a temporal cadence of 2 hours. The period from May 2010 to December 2019 is used for training with Cross-Validation and from January 2020 to April 2023 for test evaluation. The test dataset is temporally separated from the training and validation datasets, comprising samples from a distinct solar cycle. It meets the criteria given by Cinto et al. (2020) to satisfy the definition of an operational test set and allows us to assess operational performance generalisations at the beginning of a new Solar Cycle.

Each sample is associated with a photospheric line of sight magnetogram from Solar Dynamic Observatory (SDO)/HMI (Pesnell et al. (2012)) and 193Å, 211Å, 94Å, Extreme Ultraviolet (EUV) images of the solar Corona from the SDO/Atmospheric Imaging Assembly (AIA) (Lemen et al. (2012)). We selected these 3 EUV wavelengths for their strong correlation that are well suited to assemble them in an RGB-like image and leverage the full potential of pre-trained CNNs that are trained and optimised on such images. One of our objectives is also to compare the performances of photospheric magnetogram-based features and coronal thermal and morphological features to forecast Solar Flares.

3.1.1 Flare windows labels

For each sample, we compute the time-windows' labels from an extension of the Plutino et al. (2023) catalogue based on GOES's SXR flux data. To assess the regional predictive performances of the models on an operational test set, we also retrieve the positions of the flare catalogue's events for flares above the C-class threshold for samples after the 2020-01-01. To estimate the position of flares on the Sun's corona, we cross-match the flare events with images of the solar corona taken from SDO/AIA in the 171Å wavelength. The flare position tracking process starts by subtracting a coronal image at the flare's onset from the one at the flare's peak. This allows to isolate the dynamic intensity enhancement compatible with the given flare. Through the feature tracking algorithm "Trackpy" (Crocker and Grier (1996)), we can pinpoint the position of the brightest clusters of pixels and, therefore, the estimated position of the flare event in the solar corona.

3.1.2 Magnetograms

Line-of-sight magnetograms are retrieved from JSOC from the 45-second series. Original images are downsampled to 1024 by 1024 pixels with linear interpolation. The bit-depth of the pixel values is reduced from 16bit to 8bit. To maintain a good dynamic range without saturating too much, a log transform ($x \mapsto log(1+x)$) is first applied symmetrically to the positive and negative pixel values. Images are then saturated by the 99.9 percentile of the maximum pixel values over the Cross-Validation period, which corresponds to an original value of 4644G. We then apply a linear scaling so that original 0G values are centred at a pixel value of 127, with 0 and 255, respectively, representing the negative and positive saturation values.

3.1.3 EUV images

EUV AIA images are downloaded from the JSOC AIA synoptic dataset, which consists of level 1.5 AIA images available at a 2-minute cadence. The images are already registered, meaning they have a normalised arc-sec to pixel scale, one pixel being 2.4 arc-sec, and are rotated so that solar North is aligned with the y-axis. They are downscaled versions of original AIA images to a 1024 by 1024 pixel resolution. As for the magnetograms, we reduced the bit depth of the pixel values from 16 bits to 8 bits. To do so, the pixel values are first normalised by exposure-time.We then take into account AIA CCDs' degradation over time and correct for it, using the degradation correction table available from aiapy (Barnes et al. (2020)). EUV pixel distributions tend to be concentrated in a small range of values with some extreme values far from the core of the distributions, typically the results of flares or intense coronal activity. For instance, for the 193Å images, after exposure-time normalisation and instrument degradation correction, we found an average mean pixel value of 149DN with an average pixel standard deviation of 210DN, whereas the average maximum pixel value is 6247DN. While the core of the distribution is essential for retrieving morphological information on coronal structures, the extreme values might also contain important potential features associated with upcoming flares. To preserve both while maintaining a good dynamic range, we apply the same log transform and saturate each wavelength by the 99.9 percentile of their maximum pixel values over the Cross-Validation period. We then linearly scale the values between 0 and 255.

In this work, we assemble the 193Å, 211Å and 94Å 8-bit images and compress them as RGB JPEGs.

The EUV images are also processed for 1600Å, 304Å and 171Å. The resulting dataset of 1024x1024 JPEG images of magnetograms and individual wavelength EUV images is available from Francisco et al. (2024).

3.1.4 Corrupted samples

Once corrected for instrument degradation, we searched for abnormal or corrupted observations by looking for images with pixel's distribution average outside of 8 standard deviations confidence intervals computed over 48 hours running time-window. This allowed us to identify between 20 and 65 potentially corrupted observations for each channel including the magnetograms. Among them, a few corresponded to normal observations of extreme space weather events. The 94Å channel, for instance, has a very low signal-to-noise ratio outside active regions, and strong flares can punctually shift the pixel distribution average by a large amount. Besides extreme weather events, most suspicious observations appeared to correspond to problems that came from the shutter, potential repositioning of the satellite and eclipses, leaving partial to no solar disk observable in those cases. We excluded these later examples from our models' training and evaluation but left them available in the dataset along with the detailed results of this outliers study. Partially observable disks

could still be relevant in an operational context where models might still be able to forecast events coming from the visible part.

3.2 Model

3.2.1 Patch-Distributed CNN (P-CNN)

To forecast flares from full-disk images while still having an estimation of the flaring probability by regions of the solar disk, we propose to segmeent the full-disk images into patches on which we apply the same CNN. This inner-patch-CNN last layer activation is a sigmoid so that the output of each patch is bounded between 0 and 1 to represent the probability of flaring for the corresponding region. The resulting Patch-Distributed-CNN (P-CNN) aggregate the patches by taking the maximum of their outputs as the final probability and forecast at the whole disk level. We impose the inner-patch-CNN to share the same weights across the different patches to ease the convergence during training as this limits the model size and complexity. Finally, we leverage the flexibility of this architecture to remove the poles from the images as flares only emerge near the solar equator. We, therefore, crop the images at ± 614 arc-sec of latitude to obtain images that are half smaller in memory and allow for faster training. The resulting images have a resolution of 512 by 1024 pixels. In this work, we use models that received inputs downsampled at a 224 by 448 pixels resolution and patches of 112 by 112 pixels for a total of 8 patches. This additional downsampling, along with the JPEG compression, were found to no impact the resulting models performances. A summary of the resulting P-CNN architecture can be found in Figure 2.

In this study, we use the EfficientNetV2-S architecture (Tan and Le (2021)) to process the patches, with unfrozen weights pre-trained on the ImageNet dataset (J. Deng et al. (2009)) as initial weights.

The resulting models receiving magnetogram as inputs will be respectively referred as the C+ Magnetogram and M+ Magnetogram models, for the PCNN performing C+ and M+ binary forecasts. Similarly the PCNN using the 3 channel EUV combination in input will be refereed as the C+ Coronal and M+ Coronal models.

3.3 Training and Evaluation

3.3.1 Full-disk Cross-Validation Method

We train and cross-validate our models on the samples within 2010-05 and 2019-12 with a 5-fold Cross-Validation (CV). For performance evaluation to accurately reflect what would have been the performances of a model in operation, Cinto et al. (2020) highlight that a test set should include all the samples of a given period without alteration to its natural composition. We therefore keep such an operational test set, buffered by 27 days from the training and cross-validation data, ranging from 2020-01 to 2023-04, roughly the rising phase of Solar Cycle 25.

Temporal Chunking

On AR-based flare forecasts, CV-folds can be built by splitting available samples by AR numbers. This approach guarantees that samples in the training and validation sets are independent. Such independence is otherwise hard to ensure due to the important temporal autocorrelation between samples close in time. To achieve independent CV-folds in the case of full-disk forecasts, we propose a methodology based on the one proposed by Brown (2022) for solar wind forecasts. Brown (2022) constructs folds from temporal chunks of 20 days buffered by periods of 4 days, which are discarded to account for the strong autocorrelation of the solar wind speed within such periods. In the case of flares, the characteristics of a given AR can be significantly similar over a longer period of time. We, therefore, choose

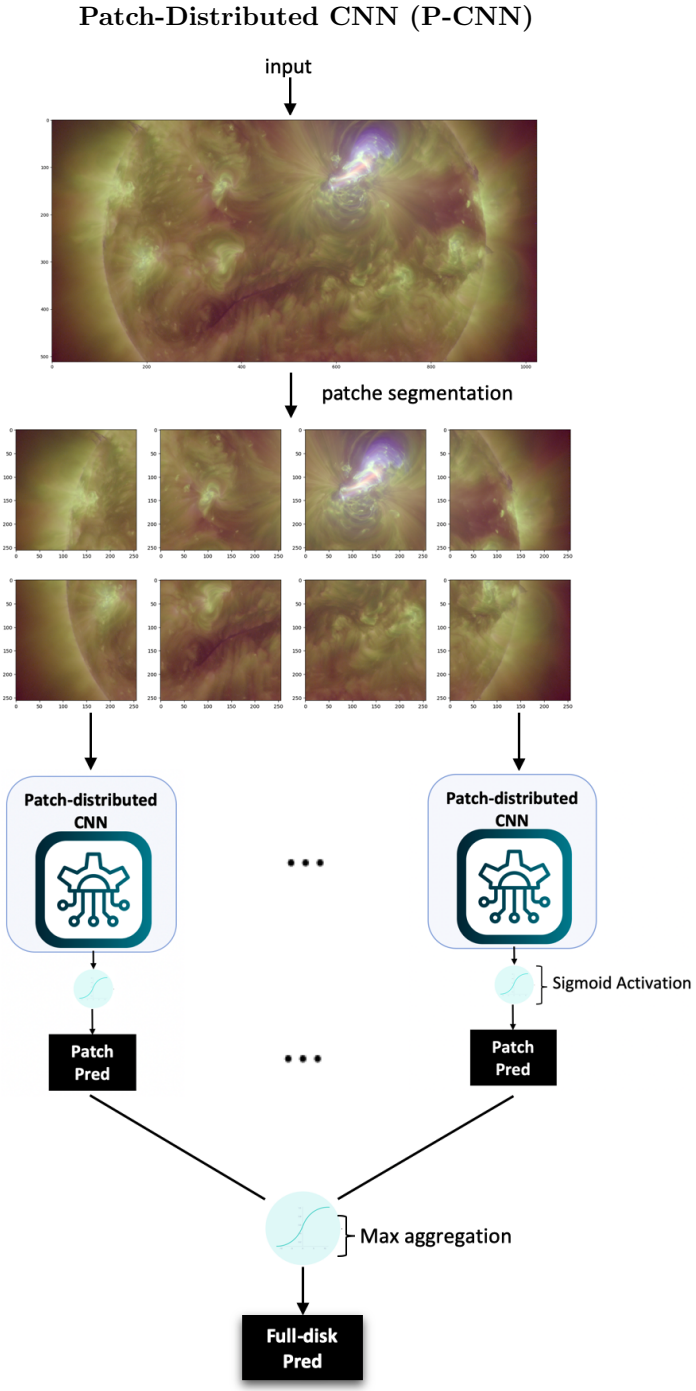


Figure 2: Architecture of the Patch-Distributed-CNN. The input is segmented into patches, each processed by a CNN that applies identical weights and outputs sigmoids representing the flaring probabilities for the corresponding regions. The patch probabilities are max-aggregated to produce the model's final output representing the flaring probability at the whole disk level, which is the target directly learned during training. In this figure, the input combines 193Å, 211Å and 94Å EUV SDO/AIA images cropped at ±614 arcsec downsampled at 224x448 pixels. The patches' size is set to 112x112 pixels, resulting in 8 patches. In this work an EfficientNet-V2-S is used as the inner-patch-CNN.

to use a buffer of 27 days, i.e. a full Carrington rotation, to ensure a strong independence between each chunk. Consequently, we choose chunks of periods of 81 days to maintain an acceptable ratio of used against discarded data. Concretely one quarter of the available data is thrown out to ensure high independence of the training and the validations sets. Unlike Brown (2022) and Pandey et al. (2023) who fill up the folds with the chunks sequentially, we allocate the chunks to obtain folds with a similar distribution of quiet, B, C, M and X class flares. It is noted that we here define a quiet time window as a window where the MPF is below the B threshold This balancing approach aims to mitigate the dataset composition's significant impact on the evaluation metrics. Guastavino et al. (2022a), for instance, demonstrated that the number of C-class flares among the non-events strongly affects the performance evaluation in the cases of M+ forecasts.

Figure 3 illustrates the CV-folds building process, with the resulting training folds depicted in blue, the validation folds in green, and the operational test set in orange.

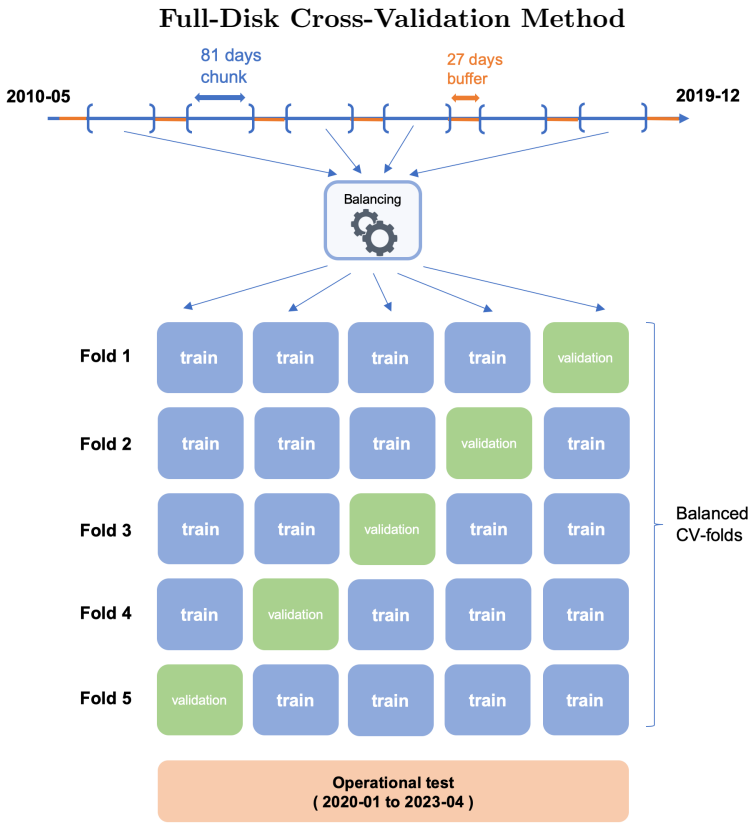


Figure 3: Temporal chunks of 81 days buffered by periods of 27 days of discarded data allow to build independent training and validation CV-folds for full-disk flare forecasting. The balancing algorithm described in section 3.3.1 allows us to allocate the chunks optimally to build folds of similar compositions. The CV-folds are built on the period ranging from 2010-05 to 2019-12. A test set, including all the samples from 2020-01 to 2023-04, chronologically split and 27 days buffered from the CV, enable to estimate what would have been the operational performances of the resulting models on the given period.

Balancing Algorithm

Due to the large chunks' size and the scarcity of certain classes, it is impossible to find a chunk allocation resulting in folds with exactly identical composition without using un-

dersampling. Consequently, we introduce an algorithm designed to achieve optimal balance, close to even compositions, before resorting to an eventual complementary undersampling. To achieve this pre-undersampling balancing, we first define the balance score b_k, defined in equation 2 :

$$b_k = \sum_{cls \in \{quiet,B,C,M,X\}} \delta_{cls} \left| \frac{n^k_{cls}}{n^*_{cls}} - 1 \right|, \tag{2}$$

where δ_{cls} is the importance-weight given to the balance of the class cls. Such quantity is introduced to account for the impossibility of achieving a perfect balance. In particular, it enables to prioritize achieving equal representation of the rarest classes before considering undersampling, as we aim to prevent further scarcity. n^k_{cls} is the number of time-windows labeled as cls in the fold k. n^*_{cls} is the targeted number of sample of class cls for every fold. It is equal to the ratio of N_{cls}, the whole number of time-windows labeled as cls within the dataset, with K, the number of folds to be built :

$$n^*_{cls} = \frac{N_{cls}}{K} \tag{3}$$

In this work we set δ_{cls} to 4 for X and quiet labelled time-windows, we set it to 2 for M labelled time-windows, and we set it to 1 for B and C labelled time-widows. The default parameters presented here provided generically satisfactory outcomes to build balanced folds with time windows ranging from 2 to 72 hours. It is noted, however, that only time windows of 24h are used in this work. To minimise the balance scores $(b_k)^K_{k=1}$, we iterate through all the available chunks to identify the chunk i and the fold k for which the decrease in b_k is maximum. This process is then repeated iteratively until all chunks are allocated. The folds' composition resulting from the balancing is shown in Figure 4 for 24h time-windows. From these folds of similar compositions, undersampling can then be used to build training and validation folds optimally aligned with specific training and validation objectives. In this study, we undersample the balanced folds to achieve a target of approximately 13,000 training samples and 2,000 validation samples for each training-validation combination. Training folds are undersampled in a way that achieves equal numbers of quiet, B, C and M time windows while retaining all available X-class samples. This training balance strategy is designed to introduce a maximum variety of subclasses, thereby encouraging our models to reach optimal performances on different sub-cases of positive and negative events. Similarly to Sun et al. (2022), we believe that balancing training sets' composition is also well-suited to train models with resulting performances less sensitive to climatological variations in operation. This might be particularly relevant in space weather applications when future climatological rates are potentially unknown and subject to changes. As for the validation folds, we undersample them in such a way as to replicate the natural climatology of the CV period. Such validation balance ensures that the performance evaluation from the validation closely mirrors how the models would have performed during the specified period. Subsequently, this allows us to assess how these performances generalize when the model is evaluated on the operational test set, eventually illustrating the impact of different climatology on the evaluation. The resulting training and validation folds are shown in Figure 5.

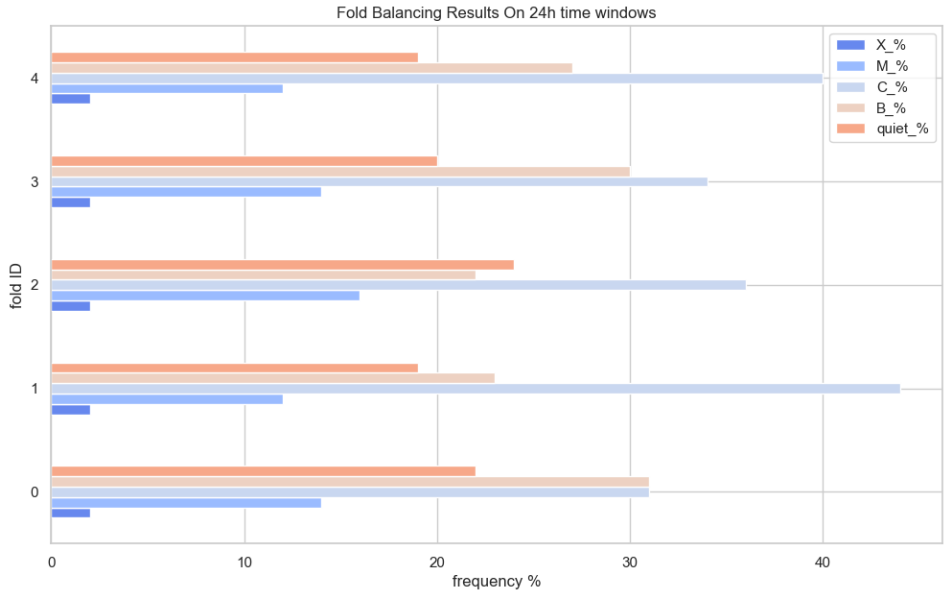


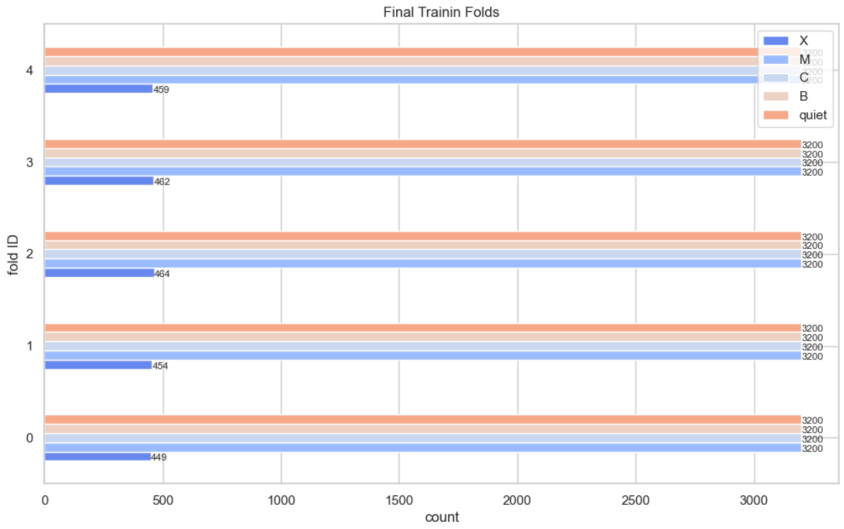
Figure 4: Folds composition resulting from the folds balancing algorithm (Section 3.3.1) for 24h time-windows

3.3.2 Models hyperparameters

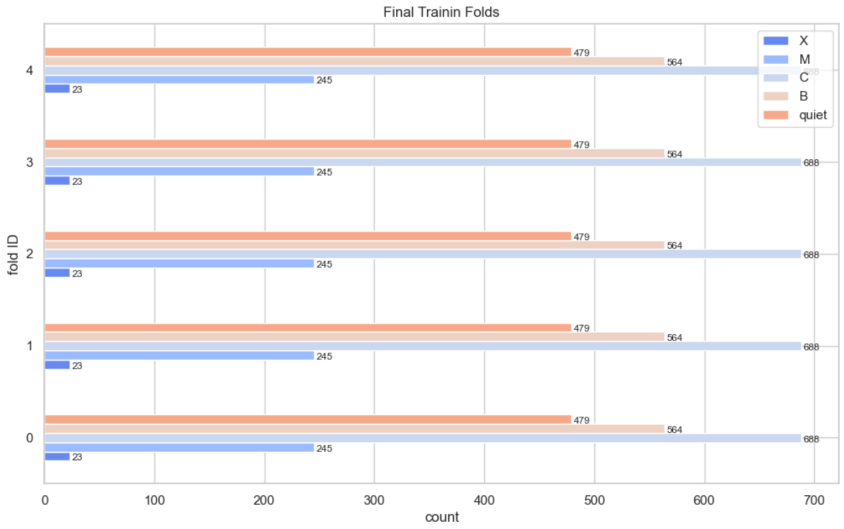
The training process is performed with TensorFlow on an Nvidia V100 GPU. The EfficientNet-V2-S used as inner-patch-CNN is initialised with pre-trained weights. Every layer but batch-normalisation ones are retrained. We use batches of 16 images and train for 15 epochs. For each fold, we save the model at the epoch with the best TSS on the validation set. We then average-ensemble the resulting folds' models as our final models when evaluating on the operational test set.

We employ the Adam optimizer (Kingma and Ba (2014)) with a decoupled weight decay regularisation (Loshchilov and Hutter (2017)). Weight decay regularisation methods consist in decaying a model's weights after every batch to provide an L2-regularisation. This helps to mitigate overfitting and foster generalisation between the training and the validation. We use a learning rate of $1e-5$ and a weight decay of $1e-4$ for every model.

We use weighted binary-cross entropy as loss functions. Despite being already close to an even balance, we weighted positive and negative events evenly in the case of the C+ models. However, for the M+ models, we over-penalize positive misclassifications with stronger weights to the the positive events with respect to the negative. This is done to help converge to solutions with higher TSS and higher recall at the expense of a worsening in the False Alarm Ratio (FAR). We do so to subsequently better highlight some limitations of the TSS in imbalanced problems. However, for more balanced models achieving optimal MCC, F1 or HSS, it is noted that such a penalisation strategy might not be recommended. Among the M+ negative sub-classes, the C-class weight is further reduced, as such events are known to be harder. This strategy encourages the model to prioritize learning what is most learnable, potentially making the gradient descent easier. The final weights that we use in the case of the M+ models are 2 for quiet time windows, 2 for B time windows, 1 for C time windows, 8 for M time windows, and 8 for X time windows.



(a) Training



(b) Validation

Figure 5: Final training and validation folds obtained from undersampling of the balanced folds from Figure 4. Validation folds replicate the climatological rates of the CV-period for better assessment of the model operational performances on such period. Training folds provide as much diversity of negative and positive subcases as possible to foster optimal performances on such subcases and optimal reliance to climatological variations. The values next to the bars indicate the exact count of the classes within each fold.

4 Results

In Section 4.1, we describe the performance metrics for full-disk observations. Notably, we highlight the stark contrast between the ostensibly state-of-the-art performance derived from conventional evaluation methodologies and the crucial limitations of the models. Section 4.2 exhibits the regional, or patch-based, performances. Subsequently, a succinct visual explanation of a prediction by the C+ Coronal model is presented in Section 4.3. Finally, Section 4.4 shows how regional predictions can be retrieved along with precise position estimations.

4.1 Full-disk-performances

4.1.1 State of the traditional performances against low Persistence-relative performances

Figure 6 and Figure 7 show the distribution of the TSS and HSS respectively for the C+ and M+ models, across validation and test sets, while the MCC, F1-score and more metrics are shown in the table 1 for the test set. The strong consistency between validation and test results indicates the effectiveness of the CV method in ensuring independence between training and validation sets. We observe that the metrics dispersion and standard deviations of the different CV-fold models are lower on the test set than the validation set. This observation suggests that the models from each fold converge to similar solutions, and that a significant share of the performance variability observed during validation may be due to dataset biases not addressed by our current fold balancing. Validation results indicate a TSS of 0.62 for the C+ magnetogram model and 0.69 for the EUV model. The M+ magnetogram yields a TSS of 0.53 for magnetograms and 0.57 for the EUV model. For the test set, we construct ensemble models where the output is the average of the probabilities produced by the individual models from each CV fold. The resulting ensembles consistently outperforms the operational performance average of the individual folds models. Their performances correspond to the stars in Figure 6 and Figure 7.

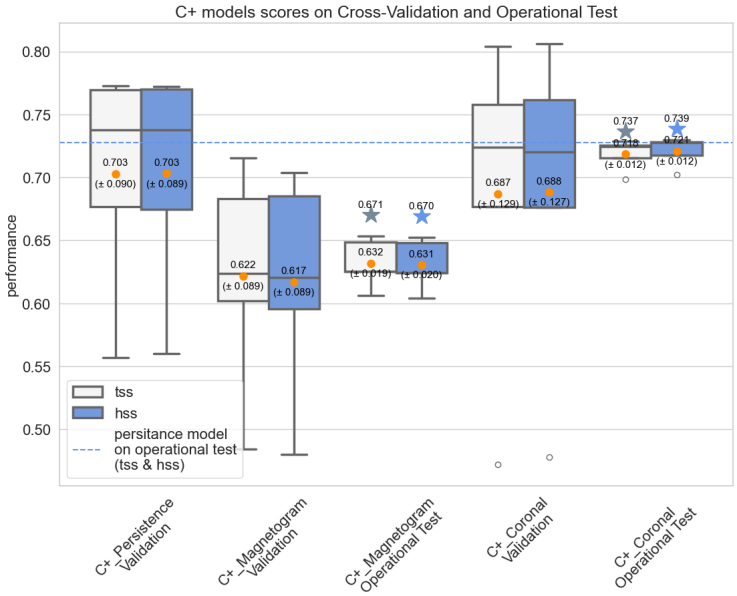
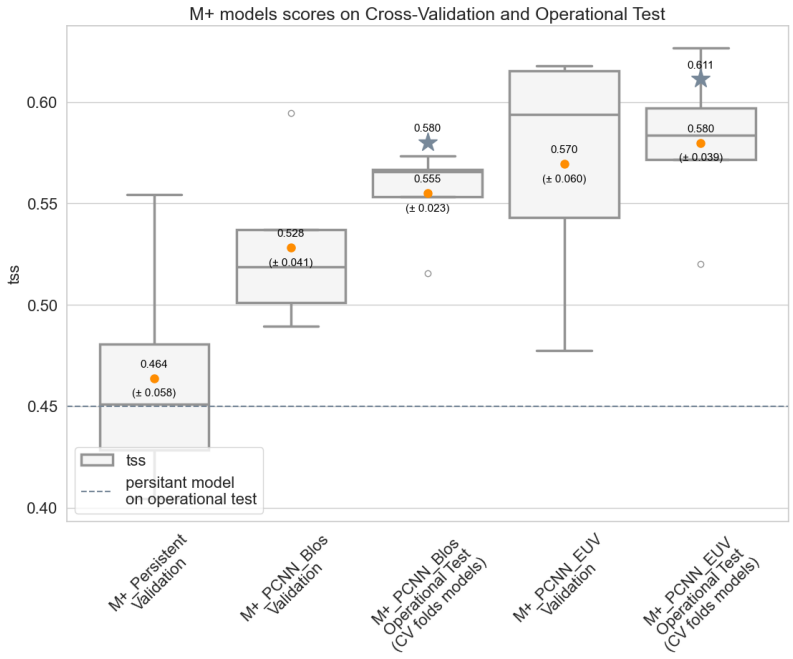
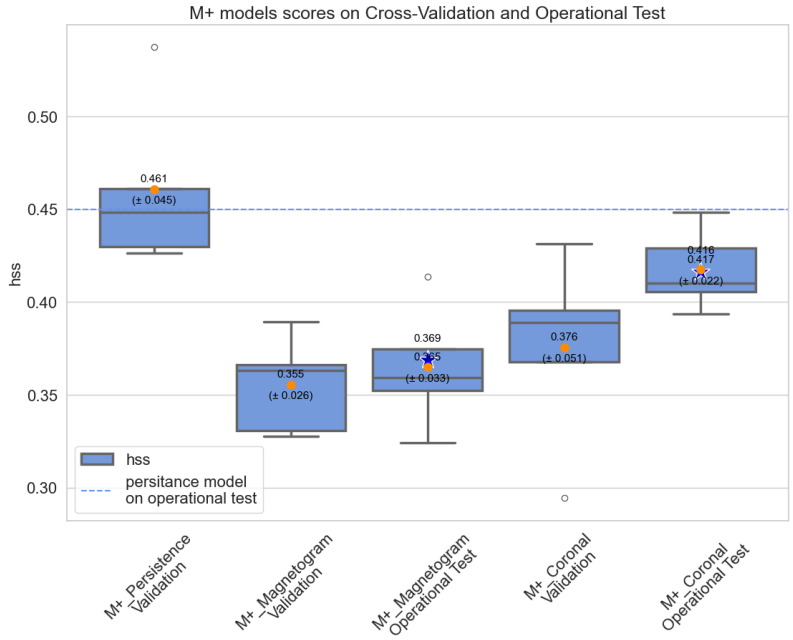


Figure 6: C+ models performances. The C+ Magnetogram model is the PCNN receiving magnetograms in input. The C+ coronal model is the PCNN receiving the 3 EUV combination images in input. The box-plots summarize the performances of the 5 models resulting from each fold. The red points and the values displayed above represent their averages. The values displayed below the red points represent one standard deviation. The stars and the values above them represent the performances of the ensemble model obtained by averaging the predictions of the 5 folds-models on the test set.

The folds-ensemble C+ models achieve an operational TSS value of 0.67 and 0.74 for Magnetogram and Coronal models, respectively. For the folds-ensemble M+ models, the operational TSS values are 0.58 and 0.61 for the Magnetogram and Coronal models, respectively. More detailed operational performances of the final ensemble models are listed in Table 1.



(a) TSS



(b) HSS

Figure 7: M+ models performances. The M+ Magnetogram model is the PCNN receiving magnetogram in input. The M+ coronal model is the PCNN receiving 3 EUV images in input. The box-plots summarize the performances of the 5 models resulting from each fold. The red points and the values displayed above represent their averages. The values displayed below the red points represent one standard deviation. The stars and the values above them represent the performances of the ensemble model obtained by averaging the predictions of the 5 folds-models on the test set.

Table 1: Full disk Performances summary on operational test set

Models	TSS	HSS	MCC	F1	Recall	FAR	ϕ	χ	PR-F1	AC-MCC	NC-MCC	NC-ϕ
C+ Persistence	0.73	0.73	0.73	0.86	0.86	0.14	0.48	0.14	0	-1	1	0.48
C+ Coronal	0.74	0.74	0.74	0.86	0.82	0.10	0.48	0.14	-0.00	0.13	0.84	0.48
C+ Magnetogram	0.67	0.67	0.68	0.82	0.77	0.11	0.51	0.14	-0.04	0.03	0.78	0.50
M+ Persistence	0.45	0.45	0.45	0.53	0.53	0.47	0.14	0.13	0	-1	1	0.08
M+ Coronal	0.61	0.42	0.46	0.53	0.82	0.61	0.14	0.13	-0.00	0.08	0.47	0.09
M+ Magnetogram	0.58	0.37	0.43	0.50	0.84	0.65	0.14	0.14	-0.06	0.06	0.43	0.09

Full disk performances of the final ensemble models on the test set. $\phi = \frac{P}{P+N}$ is the positive event ratio, $\chi = \frac{C}{P+N}$ is the activity change rate, with C the number of time-windows with an activity different from the previous one, and P and N, respectively, the number of positive and negative events. The PR-F1 (cf. Equation 1) is the F1-score Persistence-relative performance of the model. The AC-MCC (Section 2.2) assesses the models' explanatory power on time windows with an activity different from the previous one. Analogously, the NC-MCC is the MCC evaluated on time windows with the same activity as the previous one.

We put in perspective these state-of-the-art performances by considering the following points:

1. **Low Persistence Relative Performances:** The models are not significantly more explanatory than a Persistence model. Indeed, persistence models reaches in general similar state of the art performances. Despite a significant overperformance of the M+ models in terms of TSS, the lower HSS, and similar MCC and F1-scores, generally point out low persistent-relative-skills. Specifically, compared to random guesses, the models exhibit skills similar to the ones of the Persistence (similar MCC and HSS). They also show similar explanatory power (MCC) and a similar capacity to differentiate positive events from negative ones with precision (same F1-score). In particular, the null Persistent-Relative-F1 (PR-F1) (c.f. Table 1), which is the PRSS applied to the F1-score, indicates alarm systems in practice not more efficient that the Persistence models. This highlights that the increased balanced accuracy of the M+ model is mainly achieved at the expense of a worsening in precision, through an overcasting tendency favorable to the TSS on imbalanced cases. In our experiments detailed in Appendix B2, these low Persistence relative performances are found to be consistent across strong variations of the dataset composition that include different positive-event ratios and varying rates of activity change.
2. **Non-objective comparison with other works:** As discussed in Section 2.1.1 and Appendix A21, the TSS and most metrics are highly sensitive to the composition of the dataset. Consequently, it is challenging to compare the performances of this study with those from other works. However, it is noteworthy that the HSS exhibits lower sensitivity to dataset composition than the TSS, and that the HSS performances of the presented models are comparable to, or better than, those of other known models reported in the literature.

4.1.2 Performances Deficiency on Activity Change

We introduced Activity-Change performances (AC-metrics, cf. 2.2) evaluated solely on samples whose label differs from the previous consecutive non-overlapping time-window. This evaluation includes active-time-windows that follow an inactive previous window, as well as inactive-time-windows succeeding an active one (c.f. first column of Figure 1). The results reveal low AC-TSS and AC-HSS for the models (Figure 8), along with poor AC-MCC (Table 1). These findings indicate a significant lack of skill to forecast samples exhibiting changes in activity, effectively performing barely better than random guesses in such cases.

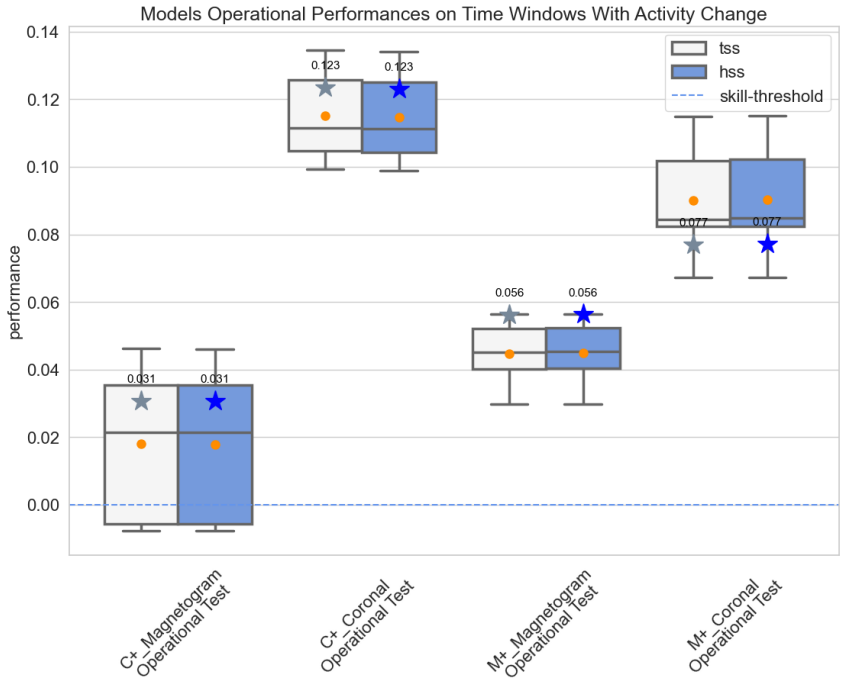


Figure 8: Models operational performances on time windows presenting an activity change compared to the previous time window. The box-plots summarize the performances of the 5 models resulting from each fold. The red points represent their averages. The stars and the values above them represent the performances of the ensemble model obtained by averaging the predictions of the 5 individual models on the test set.

4.2 Regional-performances

Table 2 summarizes the models' performances on the operational test set at a regional - or patch-level, which generally appear lower than those at the full-disk level.

These performance differences primarily stem from the significant differences in dataset composition biases. For instance, in the case of the Persistence model, it can be noted that the TSS, HSS and MCC can be expressed as a function of the the positive event ratio (ϕ) and the activity change rate (χ) (c.f. Equation A12 and Figure A1), and that these ratios are significantly less favorable at the patches level compared to the full-disk one. In particular, the positive event ratio is divided by 2.5 and 7 for the C+ and M+ models, respectively, to reach an extremely imbalanced positive event ratio of less than 2% in the M+ case. The extremely low activity change rate of 2% in the M+ case also explains the stronger degradation of the persistent-relative-performances for the M+ PCNNs, since the

deep learning models only performs better than the Persistence on AC-windows (i.e. when a change of activity occurs).

Another potential contributing factor to the performance disparity between patches and full-disk levels involves flares originating from AR spanning 2 patches. Although flares occurring near patch interstices appear well detected at the whole disk level, they can erroneously be attributed to the wrong patch based on how the explainable features are shared between the patches and the specific region of the AR from which the flare originates. The evaluation of the regional performances is thus influenced by artefacts and biases inherent to their construction. A deeper evaluation could be proposed by considering position prediction - presented in section 4.4 - matched against the actual originating AR.

Table 2: Regional Performances summary on operational test set

Models	TSS	HSS	MCC	F1	Recall	FAR	ϕ	χ	PR-F1	AC-MCC	NC-MCC	NC-ϕ
C+ Persistence	0.57	0.57	0.57	0.65	0.65	0.35	0.18	0.13	0	-1	1	0.14
C+ Coronal	0.50	0.57	0.58	0.63	0.54	0.22	0.18	0.13	-0.02	0.04	0.73	0.14
C+ Magnetogram	0.39	0.48	0.51	0.55	0.41	0.19	0.19	0.14	-0.16	0.04	0.66	0.14
M+ Persistence	0.32	0.33	0.33	0.34	0.34	0.67	0.02	0.03	0	-1	1	0.01
M+ Coronal	0.52	0.28	0.32	0.30	0.56	0.80	0.02	0.03	-0.12	0.06	0.29	0.01
M+ Magnetogram	0.43	0.22	0.26	0.25	0.48	0.83	0.02	0.03	-0.27	0.09	0.20	0.01

Regional performances of the final ensemble models on the test set. A description of the columns can be found in Table 1's caption.

4.3 Explainability

To gain insights into the coronal features learned by our models and to add positional predictions of the forecasted events, visual explainability methods are applied to the individual folds model and then averaged to represent the explainability output of the folds-ensemble model.

In particular, an analysis is conducted of the C+ Coronal P-CNN's results on 17-02-2023 at 10:00am, approximately 9 hours after the largest flare within our operational test dataset, an X2.3 flare, which started the 17-02-2023 at 18:46:52, and originated from the southern left limb of the sun. During the subsequent 24 hours, three additional regions produced C+ flares: the northern right center generated a C7.5 flare; the southern right center produced a C4.8 flare, and the northern right limb produced a C3.6 flare.

To localize the sub-region that contributed most to the patches' predictions, we use *Grad-Class Activation Maps (CAM)* (Selvaraju et al. (2016)), a generalisation of CAM (Zhou et al. (2015)) to any CNN architecture. Grad-CAM results are shown in Figure 9 and demonstrate the successful classification of each region. These results reveal that the most discriminating regions of the patches consistently encompass the positions of the forthcoming flares. The model, therefore, efficiently extracted and used features of the AR that were about to flare to make its prediction. This outcome also highlights that Grad-CAM results can help estimate predicted events' positions.

Subsequently, we employed *Guided Grad-CAM*, a technique that combines Guided Back-propagation (Springenberg et al. (2014)) with Grad-CAM, to visualise the fine-grained

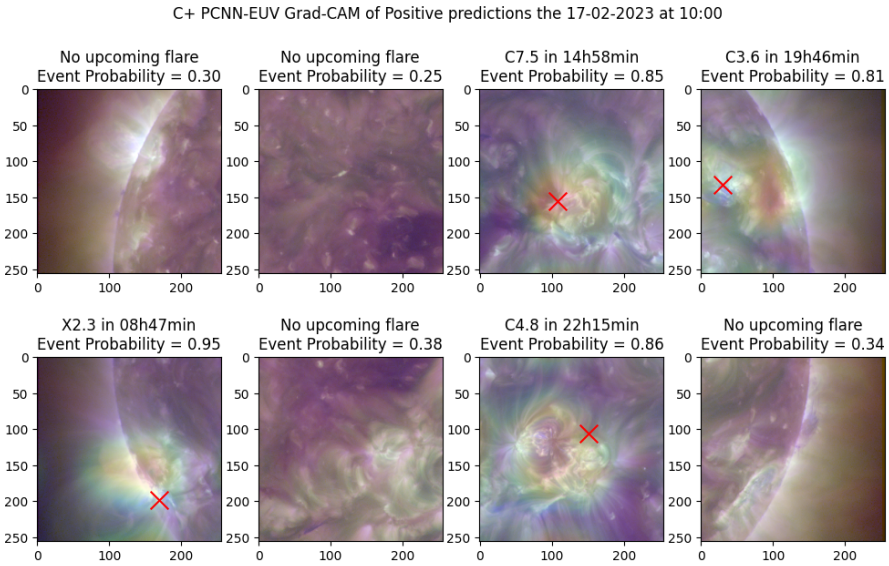


Figure 9: Grad-CAM explanation of the C+ Coronal P-CNN predictions the 17-02-2023 at 10:00am. The ground truth corresponds to the first line of the patches' titles. It is the biggest upcoming flare in the next 24 hours. The model prediction corresponds to the second line of the patches' titles. The Grad-CAM results are plotted only on positive predictions with a 'jet' color-map where the most intense values are red and the least intense are blue. The red cross represents the point where the upcoming flare will emanate accounting for rotation correction.

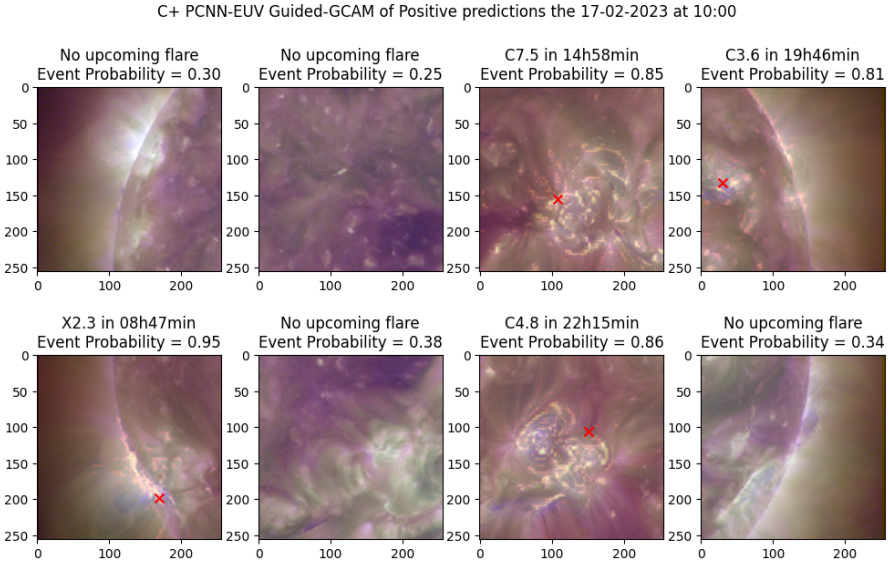


Figure 10: Guided-Grad-CAM explanation of the C+ Coronal P-CNN predictions the 17-02-2023 at 10:00am. $>(\mu + \sigma)$-masks of the guided Grad-CAM results are plotted only on positive predictions. The masks are plotted with a 'hot' color-map where the more intense the values, the more yellow and brighter.

contribution of the discriminative pixels to the model's predictions. As depicted in Figure 10, these results emphasize the model's focus on the brightest coronal structures to make its predictions. This observation aligns with the limitations of the models in predicting changes in activity. Bright coronal features are the observational signature of intense activity that correlates with highly energetic regions. This tendency supports the indication that the model primarily detects whether a region already exhibits intense energy and activity or not, rather than predicting possible changes in activity.

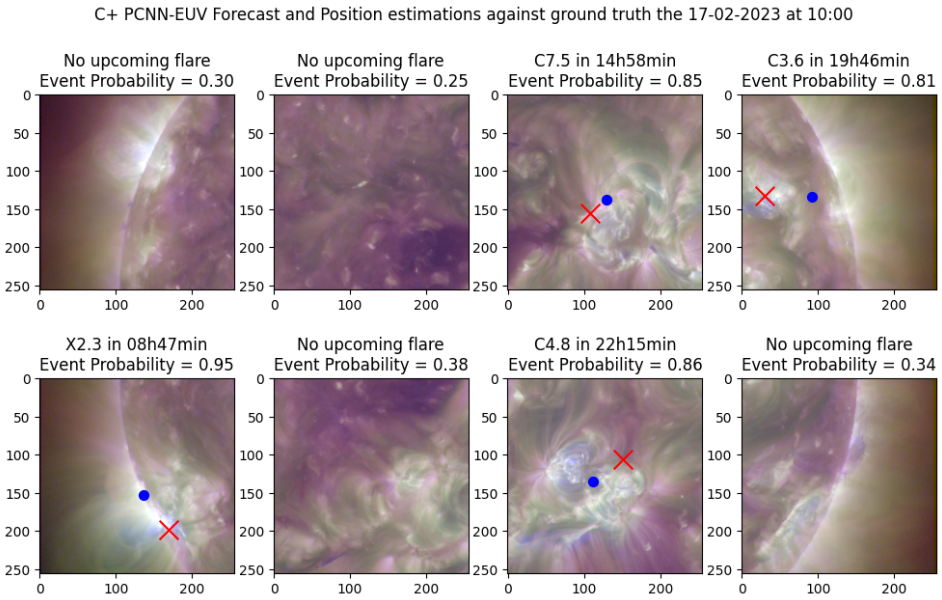


Figure 11: Example of the C+ Coronal P-CNN probabilities and positions predictions the 17-02-2023 at 10:00am. The red cross locates the point where the upcoming flare will emanate from after rotation correction. The blue point is the estimation of the upcoming flare position, calculated as the center of mass of the patch Grad-CAM results. All regions are successfully classified (classification threshold of 0.5), and the estimated positions closely align with the emergence points.

4.4 Positions predictions

We use the mass-center of the Grad-CAM results to estimate the positions of the upcoming predicted events. This information serves practical operational purposes as it can then be matched with known AR numbers, thus allowing to identify ARs likely to flare within a specific time window. The resulting model can thus potentially forecast events and identify regions prone to flaring but not yet identified or recorded in AR databases.

Figure 11 presents an illustrative example of the C+ Coronal P-CNN predictions and position estimations as of February 17, 2023, at 10:00 am. The positions' estimation accurately corresponds to the respective Active Regions for the 4 detected flares and closely align with the actual flare emergence points.

5 Discussion

5.1 Patch-distributed CNN: An Efficient Solution for Operational Models

Greater Information Retrieval with P-CNNs

The Patch Distributed CNN (P-CNN) assessed in this study offers simplicity with comprehensive insights. These models are directly trained on full-disk images and labeled accordingly, yet they retain the capability to infer event position information and estimate flare risks for sub-regions across the disk.

A potential limitation of P-CNNs arises when dealing with Active Regions spanning multiple patches. While these cases are often detected comparably well to a single full-disk model, misattribution to the wrong patch may occur. Nonetheless, this limitation can be mitigated by leveraging the precise position estimation derived from the center of mass in Grad-CAMs' results, enabling matching to known Active Regions.

Potential Regularization from P-CNNs

During the training of our most effective models, particularly those utilizing EUV images as inputs, the P-CNN demonstrated greater ease in tuning and yielded superior results compared to the use of the corresponding CNN directly on the full disk. The observed performance boost of the P-CNN over the CNN might stem from regularization effects intrinsic to the P-CNN architecture. Notably, the structural constraint of the P-CNN, confines it to a local feature space. This confinement limits the feature search to features of scales relative to the Active Regions, potentially the most pertinent scale for flare forecasts, thus easing convergence. For problems where a mix of local and large-scale features might be of interest, a global CNN layer could also be combined with the patches to achieve a sort of pyramidal features structure. Another potential regularization aspect lies in the share of the inner-CNN weight across all patches. For each sample, the inner-CNN weights must indeed converge on each patch. This cross-patches convergence artificially multiplies the number of seen instances by the number of patches. Nonetheless, confirming the regularization advantages provided by the P-CNN would necessitate a more systematic hyper-parameter grid search to confirm the generalisation of these observations and allow a deeper exploration of this hypothesis.

5.2 The current models weaknesses

Forecasting flares within 24-hour time-windows, characterized by low activity change rates, suggests a focus on metrics relative to Persistence models, which exhibit better proficiency than other no-skill models. The PR-F1 score, specifically the F1 score of the model relative to the Persistence one, is therefore of particular interest and demonstrates remarkable stability against dataset composition changes, and assesses the models' alarm system quality relatively to the Persistence. Moreover, assessments of flare forecasting metrics should distinguish between time-windows with constant and changing activity, recognizing that these scenarios represent two distinct problems. The AC-MCC and NC-MCC, which assess the models' explanatory power over activity change and constant activity time-windows, thus serve as suitable complementary performance evaluation. The PR-F1 underscores that models reaching state-of-the-art performances may still lack useful forecasting abilities compared to Persistence models, whereas the AC-MCC highlight their difficulty to forecast changes in activity better than random guesses. This might suggest that currently available features are primarily effective in identifying whether a region is already active or inactive, with no significant ability to distinguish between a region flaring yesterday and one flaring tomorrow. The question remains to identify whether some additional features could help forecast such changes efficiently. Sun et al. (2022) found an interesting performances improvement of 5 to 11% using a time-serie-based model over a point-in-time CNN. It would therefore be interesting to study the AC-performances of such models to identify a poten-

tial ability to forecast activity-changes better than random guesses. In some tests we tried including the flare history as an additional feature to a CNN, as well as using LSTM on time-series of features derived from the SXR-flux, without significant improvement with respect to a standard CNN or the Persistence models. This may suggest that the discriminative features found in magnetograms and coronal EUV images are redundant with the flare history of the previous time-window, and that the flare activity from time-window to time-window could be modeled by a Markov Process, which retains no memory beyond one single timestep. We also trained models using 8-bit uncompressed (encoded as PNG files) images at a resolution of 512x1024 pixels with the same hyperparameters and found no significant differences with the presented models. This suggests that the small-scale details lost by spatial downsampling and JPEG compression may not contain significant discriminative power. Finally, it can also be noted that there are still few studies about potential temporal coronal or chromospheric features to forecast flares. Therefore, the possibility of discovering discriminative power in such features to predict activity changes might be an interesting prospect.

Exploring novel approaches that are more aligned with the current features at our disposal could also be of interest. Magnetograms predominantly provide information about the potentially releasable energy within a specific duration. For example, the R-index (Schrijver (2009)) is related to the magnetic energy concentration around the Polarity Inversion Line (PIL)s, where instability is most likely to occur. Therefore, it might be more feasible to forecast quantities more closely related to the energy emitted by flares during specific time windows. For instance, labeling time-windows based on the sum of the SXR-fluence - i.e., the integrated flux - produced by flares might better characterise the activity occurring during these periods. This new characterization would also be a more reliable proxy of the released energy and could be more explainable. Assessing the emitted SXR-fluence within a time-window could also offer valuable insights for astronauts.

5.2.1 Activity change definition limit

Our current definition of AC and NC windows has some limitations. An AR about to flare and emerging from the far side of the Sun, where it would have already flared in the previous time-window, would physically represent a NC-windows that is constantly active. However, it could lead to a labeling of the current time-window as an AC-window, changing from inactive to active, if the visible disk was previously inactive. A similar case may arise at the patch level when an AR crosses the boundaries between two patches. Furthermore, the definition of changing and constant relies on the arbitrary length of our time-windows. For a 24-hour time window, an AR flaring at 25h of intervals will represent (AC) time-windows while an AR flaring at 23h of intervals can represent (NC) time-windows (here constantly active), whereas the difference between the two cases might not be meaningful physically. Such artefacts could be mitigated by using temporally weighted metrics as introduced by Guastavino et al. (2022). However, it is worth noting that the previous limitations mostly contributes positively to the AC-performances. Indeed, these limitations cause some time-window physically closer to a constant activity to be labeled as a change of activity. These constant activities, which are more explainable than true activity changes, might therefore slightly boost the AC performances. Hence, the explanatory power of the proposed model on AC-windows could in fact be even lower than assessed with the presented AC performances.

6 Conclusion

Our study introduces a novel method to facilitate the construction of balanced and independent CV-folds for full-disk flare forecasting with minimal undersampling. Introducing P-CNNs trained with this method, our models achieve state-of-the-art performances at full-disk levels on the rising phase of Solar Cycle 25, with a TSS of 0.74 for C+ forecasts and 0.62 for M+ forecasts using EUV coronal images as inputs, revealing an interesting performance edge over the magnetograms. The Patch-CNN architecture offer the advantage of training models on full-disk images using full-disk labels alone while providing forecasts for sub-regions of the disk and estimations of events' position. This results in minimal operational external dependencies and the reduction of mislabelling issues during training and evaluation.

Our discussion underscores the challenge of using conventional metrics in flare forecasting, as they are significantly sensitive to various elements of datasets' composition and do not enable identifying important model limitations. Notably, classical metrics may hinder objective model comparisons across different works and optimal model selection during training for operational purposes.

To address these issues, we introduced Persistence-Relative-Skill-Scores along with evaluation restrictions to time-windows exhibiting changes of activities (AC-performances) and those without changes of activity (NC-performances). These metrics are well-suited to the specific challenges of flare forecasting and can help improve model comparisons.

The PR-F1 and AC-MCC reveal low forecasting skills when comparing our models to Persistence ones or random models for time-windows with changes in activity. Both the PR-F1 and MCC exhibit greater stability and reliability than the commonly used HSS, the latter being already more reliable and informative than the TSS in imbalanced cases.

Appendix A Metrics complements

A1 Formulas

A11 Confusion Matrix

Evaluation metrics for binary classification are defined as a function of the confusion matrix CM:

$$CM = \begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix} \tag{A1}$$

where True Positive (TP) and True Negative (TN) denote the number of positive and negative events correctly classified, while False Positive (FP) and False Negative (FN) are the number of misclassified ones.

A12 Basic Confusion Matrix

The following rates summarize the information contained in the confusion matrix :

- Class-Accuracy rates

$$\text{Recall} = \text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{A2}$$

$$\text{Sensitivity} = \text{True Negative Rate (TNR)} = \frac{TN}{TN + FP} \tag{A3}$$

- Class-Precision rates:

$$\text{Negative Predictive Value (NPV)} = \frac{TN}{TN + FN} \tag{A4}$$

$$\text{Precision} = \text{Positive Predicitve Value (PPV)} = \frac{TP}{TP + FP} \tag{A5}$$

  Equivalently to the precision, practitioners alternatively look at the FAR which is the complementary of the former and gives the rate at which positive predictions give a false alarm.

A13 F1-Score

The $F1$-score, defined in equation A6, is the harmonic mean of the precision and the recall.

$$F1 = 2 * \frac{precision * recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{A6}$$

It offers a consolidated assessment of an alarm system skill when the emphasis lies on achieving high recall (TPR) to detect maximum events, along with high precision (PPV) to ensure confidence in positive predictions.

A14 True Skill Statistic (TSS) / Informdness

The True Skill Statistic (TSS) was introduced to evaluate weather forecasts by Hanssen and Kuipers (1965). In other fields, it is also known as the (bookmaker) informedness, Peirce's index or Younden's J index. It can be dated back to 1884 (Peirce (1884)). It is equal to the difference between the true positive rate and the false positive rate but also

to the balanced accuracy re-scaled between -1 and 1, i.e. the average of the class-accuracy rates (TPR and TNR) normalized in [-1,1]:

$$TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN} = TPR + TNR - 1 \tag{A7}$$

Random and constant models produce a TSS of 0.

A15 Markdness (MK)

The markedness is the precision equivalent of the TSS; it is the average of the class-precision rates (PPV and NPV) normalized in [-1,1].

$$MK = PPV + NPV - 1 \tag{A8}$$

A16 Negative Frequency Bias (NFB)

$$NFB = \frac{TN+FN}{TN+FP} = \frac{\text{Predicted Neagtives}}{\text{Observed Negatives}} \tag{A9}$$

A17 Heidke Skill Score (HSS) / Cohen's Kappa index

The HSS, defined in equation A10, was introduced to evaluate weather forecasts by Heidke (1926). In other fields, it can be known as Cohen's Kappa index. It is commonly used in flare forecasting to compare a model skill relatively to a random guess model (Camporeale (2019)).

$$HSS = 2 * \frac{TP*TN - FN*FP}{P(TN+FN) + N(TP+FP)} \tag{A10}$$

The HSS varies between -1 and 1, with 1 denoting the performance of a perfect classifier and 0 indicating the one of random guesses. It can then be noted that the HSS is the harmonic mean between $\frac{TSS}{NFB}$ and $MK * NFB$ (see Delgado and Tibau (2019) for mathematical proof). The HSS is, therefore, a weighted harmonic average between the TSS and the MK, with a model-dependent importance given to each.

A18 Matthews Correlation Coefficient (MCC)

The MCC was introduced by Matthews (1975) to address class imbalances in performance evaluation. It is the Pearson correlation coefficient between binary predictions and labels :

$$MCC = \frac{TP*TN - FN*FP}{\sqrt{P(TN+FN)*N*(TP+FP)}} \tag{A11}$$

The MCC ranges between -1 and 1. Similar to the TSS and HSS, both random and constant models produce an MCC score of 0. The MCC is the geometric average between the TSS and the Markdness (c.f. Delgado and Tibau (2019) and Chicco, Tötsch, and Jurman (2021)). It thus summarizes the four basic confusion matrix rates with equal weights given to each.

A2 Metrics' Notable Properties

A21 The TSS sensitivity to the dataset composition

Bloomfield et al. (2012) proposed the TSS for flare forecasting as, in simplified cases, it is found to be insensitive to the class-balance. This has been argued to make it a suitable metric for comparing models among different datasets with varying class balances (Bloomfield et al. (2012), Chicco, Tötsch, and Jurman (2021)). However, in the case of flare forecasting, we prove that usual models will have a TSS that is strongly sensitive to the positive events ratio. Indeed, the mathematical independence between the TSS and the class balance only holds for models that perform equally in every possible case of negative and positive events. For flare forecasting models, the weak performances on samples exhibiting changes in activity result in a direct non-linear dependency of the TSS to the positive event ratio and, thereby, the class balance.

Let us consider the limit case of a model that perfectly identifies time-windows without activity change but consistently fails when activity change occur. Such a model can be known as a Persistence model. On a given time period, if evaluated on every time-window, the Persistence model's number of FP will be equal to the number of FN which will be equal to half the number of activity change $\frac{C}{2}$. The number of TP will be $P - \frac{C}{2}$, and the number of TN will be $N - \frac{C}{2}$, where P and N are respectively the number of positive and negative events. If we denote $\phi = \frac{P}{P+N}$ the positive event ratio, and $\chi = \frac{C}{P+N}$ the activity change rate, the TSS, HSS and the MCC from equation A7, A10 and A11 simplify in :

$$TSS_{Persistence} = HSS_{Persistence} = MCC_{Persistence} = 1 - \frac{1}{2}\frac{\chi}{\phi(1-\phi)} \tag{A12}$$

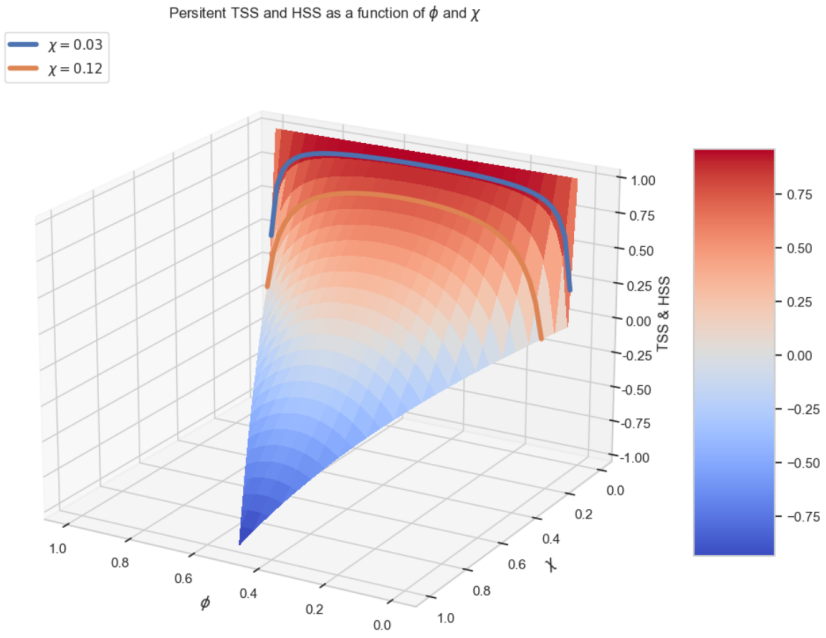
We may additionally note that as C is at maximum equal to twice the minimum between P and N. Therefore, the equation A12 is defined only when :

$$\chi \leq \begin{cases} 2\phi & \text{if } \phi \leq 0.5 \\ 2(1-\phi) & \text{if } \phi \geq 0.5 \end{cases} \quad \text{with } \phi \in [0,1] \tag{A13}$$

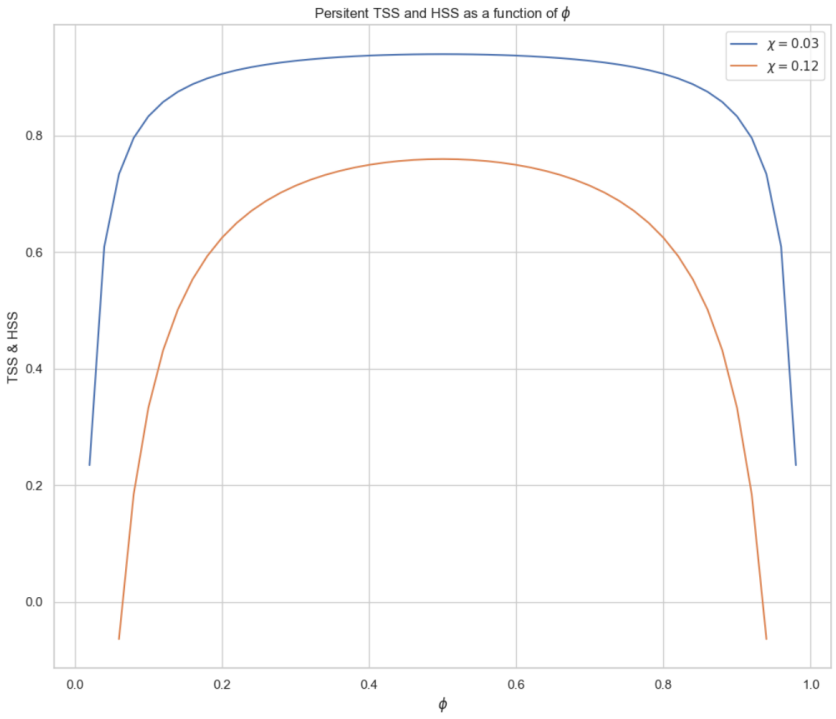
Figure A1 shows the plot of a Persistence model's TSS, HSS and MCC according to equations A12 and A13. The metric linearly increase with the decrease of the activity change rate χ which is typically low in flare forecasting. The performance is non-linearly dependent on the positive event ratio ϕ with a stronger sensitivity in most imbalanced cases. For a variation of ϕ from 6 to 12%, the TSS increases from -0.06 to 0.44 with a standard activity change rate of 0.12. The model is therefore deemed unskilled in the first case, whereas it can be estimated as mildly proficient in the second case, only because of a doubling of the positive event ratio.

In practice flare forecasting models can be expected to have a similar TSS, HSS and MCC sensitivity to ϕ and χ as they have good performances on time windows with the same activity as the previous one, whereas they struggle on time windows with changing activity with respect to the previous one.

The activity change rate χ and the positive event ratio ϕ thus emerge as fundamental dataset biases in flare forecasting. Our experiment in Appendix B, show that empirically the sensitivity of the models' metrics to these biases is actually stronger than the one of the Persistence models. In particular, the TSS seems to exhibit heightened sensitivity to dataset biases compared to the HSS, and the MCC appears as the most stable of the three. This is likely due to the positive influence of overcasting on TSS, as well as additional models' flaws, such as the low accuracy on C-class negative events in M+ forecasts. These factors further complexify the sensitivity of the models' performance to various dataset biases.



(a) The TSS & HSS dependency to ϕ and χ for persistent models



(b) The TSS & HSS non-linear dependency to ϕ for fixed χ on persistent models

Figure A1: Figure (a) displays the plot of a Persistence model's TSS, HSS and MCC as a function of ϕ and χ. Figure (b) displays the variations of the TSS, HSS and MCC with ϕ for two different constant χ. $\phi = \frac{P}{P+N}$ is the positive event ratio. $\chi = \frac{C}{P+N}$ is the activity change rate

A22 TSS's incomplete information for operational model purposes

The TSS is inadequately informative in highly imbalanced scenarios, where a good TSS might obscure strong over- or undercasting tendencies, as highlighted by Leka et al. (2019). While a high TSS ensures accurate classification for most positive and negative events, it does not always translate into a practical model intended for an alarm system if the evaluation is made on an imbalanced set. In cases of substantial imbalance, a high TSS can be reached with a precision close to 0, resulting in a false alarm ratio close to 1, rendering the model unfit for an alert system. To illustrate, let us compare the two following models using the same synthetic dataset with a positive event ratio of 0.001:

Model 1
TP = 99, FN = 1, FP = 5 000 and TN = 94 900, then,
TSS = 0.94, recall = 0.99, precision = 0.02, f1 = 0.04.
A case similar to this one might arise with X+ flare binary classifiers. A comparative example can be found with the X+ flares binary classifier of Huang et al. (2018) achieving a TSS of 0.714 for a False Alarm Ratio of 0.98.

Model 2
TP = 90, FN = 10, FP = 50 and TN = 94 851, then,
TSS = 0.90, recall = 0.90 precision = 0.64, f1 = 0.75.
This model, while having a lower TSS, maintains a reasonably good recall with significantly higher precision, making it arguably preferable over model 1 for operational purposes.

The TSS contains no information about a model's precision and is, therefore, not a well-suited indicator to select a model for an alarm system in imbalanced cases. Without preferences defined between recall and precision, the F1 score proves more informative in discriminating between a useful or impractical model in operation. Specific recall and precision preferences can also be considered using the $F\beta$ score, which extends the F1 score by giving β times more importance to the recall than the precision.

A23 HSS interpretations

The HSS is a weighted harmonic average between the TSS and the MK, with model-dependent importance given to each. The HSS thus synthesizes information both about a model's accuracy and its precision in the different classes, making it arguably a more suitable metric for assessing a model's suitability as an alarm system than the TSS. However, the model-dependent weight importance between the Markedness and the TSS makes it complex to interpret and compare models. The harmonic mean mathematically gives more importance to smaller values. Consequently, for a model tending to undercast the negative class, i.e. NFB smaller than 1, the contribution of the Markedness to the HSS is increased. For a model tending to overcast the negative class, i.e. NFB larger than 1, the importance of the TSS contribution to the HSS is conversely increased.

A24 MCC advantages over other metrics

While the MCC is still uncommon in Space Weather, it is argued, for general cases, to be a more informative and reliable metric compared to the accuracy and the F1 score (Chicco and Jurman (2020)), the TSS (Chicco, Tötsch, and Jurman (2021)), the HSS (Delgado and Tibau (2019),Chicco, Warrens, and Jurman (2021)). The MCC is also to be favored over other metrics such as the ROC AUC (Chicco and Jurman (2023)) and the Brier Score (Chicco, Warrens, and Jurman (2021)), two other metrics of interest in flare forecasts. Empirically, we showed (c.f. Appendix B) the MCC scores to be more resilient to dataset composition changes compared to the HSS, which, in turn, is more stable than the TSS. Consequently, the MCC might be preferable for both model selection and comparison across different datasets. It is often a better choice compared to the HSS, as it shares similar infor-

mation but demonstrates higher stability in extreme cases and is a consistent synthesis of models' class accuracies and precisions. The MCC also allows to measure models explanatory power agnostically of users preferences. Despite the MCC's comprehensive assessment of a model's overall quality, the F1-score remains relevant due to its straightforward interpretability for operational alarm system applications. The choice of one metric among the others should ultimately be decided by the importance given to each class and their accuracy and precision. (Chicco, Tötsch, & Jurman, 2021) summarizes that F1 might still be preferred over the MCC when the accurate and confident classification of positive elements holds greater importance than for negative ones. The TSS, on the other end, is still relevant on balanced problem, or on imbalanced problem if no importance is given to the models precision.

Appendix B Empirical Variability Of Standard Metrics to the Evaluation Set Composition

B1 Standard Metrics

Metrics can be linked with the positive event ratio and the activity change rate in a non-linear way. For instance, the recall and the F1-score of a Persistence model is proportional to their ratio : $F1_{persistant} = 1 - \frac{\chi}{2\phi}$. In appendix A21 and Figure A1 we exposed the more complex relationship of the TSS and HSS of a Persistence model with these ratios. Similar bias sensitivity should be expected for every model with significant skills deficiency on activity changes. To observe empirically the impact of the two ratios on the metrics evaluated on our models, we display the model's performance variation for different combinations of those ratios in Figure B1. The sub-test samples with varying compositions are obtained by varying the start of the test set from 2020-01-01 to 2023-01-01 while maintaining 2023-04-18 as the end date.

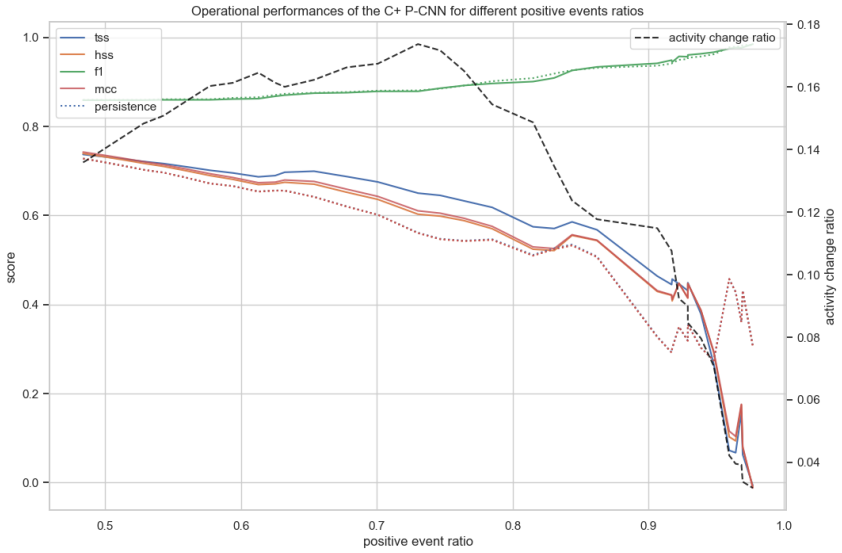
All the metrics appear strongly sensitive to the dataset composition. The TSS appears to be the most affected, especially in the imbalanced case of the M+ models, while the F1-score and the MCC are the most stable ones. It is worth noting that the F1-score is defined on the range [0, 1], while the Matthews correlation coefficient (MCC) is defined on the range [-1, 1]. This implies that a unit change in the F1-score corresponds to a double change in magnitude compared to the MCC relative to their respective definition intervals.

B2 Persistence relative skill scores

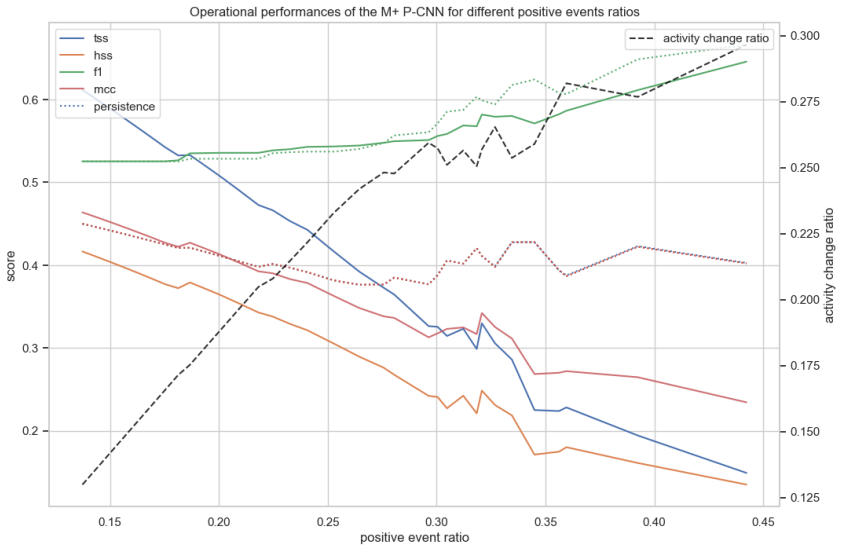
Using the same dataset composition variations as presented in the previous section (Appendix Appendix B) the variability of the Persistence-relative-skill-scores are displayed in Figure B2.

The Persistence relative skill scores appear to vary less than their standard metric equivalent. The most resilient ones appear to be the PR-F1 followed by the PR-MCC. The PR-F1 in particular appears remarkably stable except with the C+ model in the most extreme class-imbalances, where it becomes slightly positive.

With the exception of the PR-TSS, the Persistence relative skill scores indicate a consistent lack of performance in the M+ case, and null to slightly positive for the C+ model. This suggests that despite the strong impact of the dataset biases on the performance evaluation, models reaching state-of-the-art performances could be expected to consistently struggle to outperform Persistence models over varying subsets of the Solar Cycle.

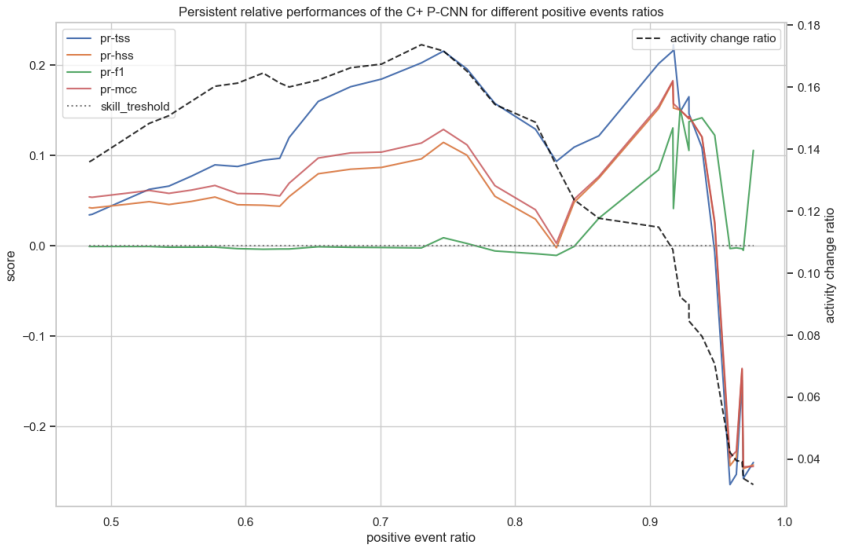


(a) TSS

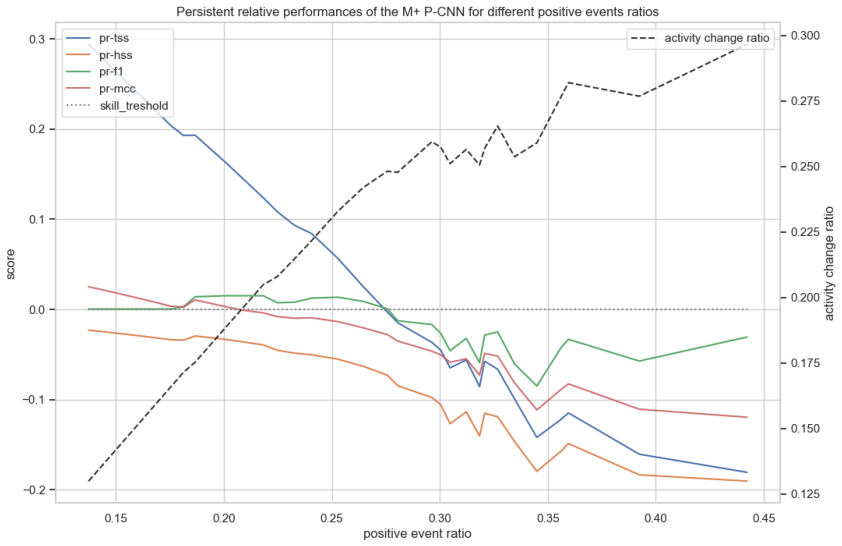


(b) HSS

Figure B1: EUV models operational performances against variations of the positive event ratio and the activity change rate. The sub-test-sets with varying compositions are obtained by sliding the start of the test set from 2020-01-01 to 2023-01-01. The left Y-axis represent the metrics score. The X-axis are the varying positive event ratios. The right Y-axis represents the activity change rates which are plotted as a black dashed line.



(a) TSS



(b) HSS

Figure B2: EUV models operational Persistence-relative-performances against variations of the positive event ratio and the activity change rate. The axis and sub-test-sets are the same as in Figure B1

Appendix C List of Acronyms

AR	Active Region
CAM	Class Activation Maps
CNN	Convolutional Neural Network
EUV	Extreme Ultraviolet
FAR	False Alarm Ratio
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
PIL	Polarity Inversion Line
SDO	Solar Dynamic Observatory
TSS	True Skill Statistic
PRSS	Persistent Relative Skill Score
PR-F1	Persistent-Relative-F1
HSS	Heidke Skill Score
MCC	Matthews Correlation Coefficient
P-CNN	Patch-Distributed-CNN
TPR	True Positive Rate
TNR	True Negative Rate
PPV	Positive Predictive Value
NPV	Negative Predictive Value
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
AIA	Atmospheric Imaging Assembly
SXR	Soft X-Rays
MPF	Maximum Peak Flux
CV	Cross-Validation
NFB	Negative-Frequency-Bias

Appendix D Open Research

The data prepared for this study offers a compact and Machine-learning-ready dataset that can be used for other applications and is available from: Francisco et al. (2024). It is derived from the AIA-Synoptic dataset `http://jsoc.stanford.edu/data/aia/synoptic/` and JSOC's level 1.5 45 seconds series HMI LOS magnetograms. The time-window labels are derived from an extension (Plutino et al. (2024)) of the Plutino's flare event catalogue (Plutino et al. (2023)).

The code - based on TensorFlow (Abadi et al. (2015)) - developed to analyse and train the models of this work are also publicly available and notebooks are provided to replicate the results presented in this study: `https://github.com/gfrancisco20/flare_limits_pcnn`

Acknowledgments

This research is part of the SWATNet project which is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant Agreement No 955620.

This research has also been carried out in the framework of the CAESAR project, supported by the Italian Space Agency and the National Institute of Astrophysics through the ASI-INAF n.2020-35-HH.0 agreement for the development of the ASPIS prototype of the scientific data centre for Space Weather.

This study was also produced within the IA and the CITEUC. IA is supported by Fundação para a Ciência e a Tecnologia (FCT, Portugal) through the research grants UIDB/04434/2020 and UIDP/04434/2020. CITEUC, is funded by National Funds through FCT - project UIDP+UIDB/00611/2019.

Michele Berretti acknowledges that this publication (communication/thesis/article, etc.) was produced while attending the PhD program in Space Science and Technology at the University of Trento, Cycle XXXIX, with the support of a scholarship financed by the Ministerial Decree no. 118 of 2nd March 2023, based on the NRRP - funded by the European Union - NextGenerationEU - Mission 4 "Education and Research", Component 1 "Enhancement of the offer of educational services: from nurseries to universities" - Investment 4.1 "Extension of the number of research doctorates and innovative doctorates for public administration and cultural heritage"

Project partially funded under the National Recovery and Resilience Plan (PNRR), Missione 4 "Istruzione e Ricerca" – Componente C2 – Investimento 1.1, "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)" – Call for tender No. 1409 of 14/09/2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU Award Number: P2022RKXH9, Concession Decree No. 1397 of 06/09/2023 adopted by the Italian Ministry of University and Research, project CORonal mass ejection, solar eNERgetic particle and flare forecaSTing from phOtospheric sigNaturEs (CORNERSTONE).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `https://www.tensorflow.org/` (Software available from tensorflow.org)

Barnes, G., Leka, K. D., Schrijver, C. J., Colak, T., Qahwaji, R., Ashamari, O. W., ... Wagner, E. L. (2016, October). A Comparison of Flare Forecasting Methods. I. Results from the "All-Clear" Workshop. *Astrophys. J.*, *829*(2), 89. doi: 10.3847/0004-637X/829/2/89

Barnes, W. T., Cheung, M. C. M., Bobra, M. G., Boerner, P. F., Chintzoglou, G., Leonard, D., ... Wright, P. J. (2020). aiapy: A python package for analyzing solar euv image data from aia. *Journal of Open Source Software*, *5*(55), 2801. Retrieved from https://doi.org/10.21105/joss.02801 doi: 10.21105/joss.02801

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

Brown, E. J. E. (2022, March). Attention-based machine vision models and techniques for solar wind speed forecasting using solar EUV images. In *Proceedings of the 2nd machine learning in heliophysics* (p. 47).

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166-1207. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018SW002061 doi: https://doi.org/10.1029/2018SW002061

Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. Retrieved from https://doi.org/10.1186/s12864-019-6413-7 doi: 10.1186/s12864-019-6413-7

Chicco, D., & Jurman, G. (2023). The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. Retrieved from https://doi.org/10.1186/s13040-023-00322-4 doi: 10.1186/s13040-023-00322-4

Chicco, D., Tötsch, N., & Jurman, G. (2021). The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, *14*(1), 13. Retrieved from https://doi.org/10.1186/s13040-021-00244-z doi: 10.1186/s13040-021-00244-z

Chicco, D., Warrens, M., & Jurman, G. (2021, 05). The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2021.3084050

Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. (2020, July). A framework for designing and evaluating solar flare forecasting systems. *Mon. Not. Roy. Astron. Soc.*, *495*(3), 3332-3349. doi: 10.1093/mnras/staa1257

Crocker, J. C., & Grier, D. G. (1996). Methods of digital video microscopy for colloidal studies. *Journal of Colloid and Interface Science*, *179*(1), 298-310. Retrieved from https://www.sciencedirect.com/science/article/pii/S0021979796902179 doi: https://doi.org/10.1006/jcis.1996.0217

Delgado, R., & Tibau, X.-A. (2019). Why cohen's kappa should be avoided as performance measure in classification. *PLoS One*, *14*(9), e0222916. doi: 10.1371/journal.pone.0222916

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Deng, Z., Wang, F., Deng, H., Tan, L., Deng, L., & Feng, S. (2021, dec). Fine-grained solar flare forecasting based on the hybrid convolutional neural networks. *The Astrophysical Journal*, *922*(2), 232. Retrieved from https://doi.org/10.3847/1538-4357/ac2b2b doi: 10.3847/1538-4357/ac2b2b

Deshmukh, V., Flyer, N., van der Sande, K., & Berger, T. (2022, May). Decreasing False-alarm Rates in CNN-based Solar Flare Prediction Using SDO/HMI Data. *Astrophys. J. Suppl.*, *260*(1), 9. doi: 10.3847/1538-4365/ac5b0c

Francisco, G., Del Moro, D., Barata, T., & Fernandes, J. (2024, April). *Sdo 2h machine learning dataset.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.11058938 doi: 10.5281/zenodo.11058938

Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. (2022a, June). Implementation paradigm for supervised flare forecasting studies: A deep learning application with video data. *Astron. Astrophys.*, *662*, A105. doi: 10.1051/0004-6361/

202243617

Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. (2022b, September). Operational solar flare forecasting via video-based deep learning. *arXiv e-prints*, arXiv:2209.05128. doi: 10.48550/arXiv.2209.05128

Guastavino, S., Piana, M., & Benvenuto, F. (2022). Bad and good errors: value-weighted skill scores in deep ensemble learning. *IEEE transactions on neural networks and learning systems*. doi: 10.1109/TNNLS.2022.3186068

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

Heidke, P. (1926). Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska Annaler*, *8*(4), 301–349. doi: 10.1080/20014422.1926.11881138

Huang, X., Wang, H., Xu, L., Liu, J., Li, R., & Dai, X. (2018, March). Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms. *Astrophys. J.*, *856*(1), 7. doi: 10.3847/1538-4357/aaae00

Kingma, D. P., & Ba, J. (2014, December). Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980

Leka, K. D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., ... Terkildsen, M. (2019, August). A Comparison of Flare Forecasting Methods. II. Benchmarks, Metrics, and Performance Results for Operational Solar Flare Forecasting Systems. *Astrophys. J. Suppl.*, *243*(2), 36. doi: 10.3847/1538-4365/ab2e12

Lemen, J. R., Title, A. M., Akin, D. J., Boerner, P. F., Chou, C., Drake, J. F., ... Waltham, N. (2012, January). The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *Solar Phys.*, *275*(1-2), 17-40. doi: 10.1007/s11207-011-9776-8

Li, X., Zheng, Y., Wang, X., & Wang, L. (2020, March). Predicting Solar Flares Using a Novel Deep Convolutional Neural Network. *Astrophys. J.*, *891*(1), 10. doi: 10.3847/1538-4357/ab6d04

Loshchilov, I., & Hutter, F. (2017, November). Decoupled Weight Decay Regularization. *arXiv e-prints*, arXiv:1711.05101. doi: 10.48550/arXiv.1711.05101

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442-451. Retrieved from `https://www.sciencedirect.com/science/article/pii/0005279575901099` doi: https://doi.org/10.1016/0005-2795(75)90109-9

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. (2018, May). Deep Flare Net (DeFN) Model for Solar Flare Prediction. *Astrophys. J.*, *858*(2), 113. doi: 10.3847/1538-4357/aab9a7

Pandey, C., Angryk, R. A., & Aydin, B. (2023, July). Explaining Full-disk Deep Learning Model for Solar Flare Prediction using Attribution Methods. *arXiv e-prints*, arXiv:2307.15878. doi: 10.48550/arXiv.2307.15878

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., & Shin, G. (2018, December). Application of the Deep Convolutional Neural Network to the Forecast of Solar Flare Occurrence Using Full-disk Solar Magnetograms. *Astrophys. J.*, *869*(2), 91. doi: 10.3847/1538-4357/aaed40

Peirce, C. S. (1884, November). The Numerical Measure of the Success of Predictions. *Science*, *4*(93), 453-454. doi: 10.1126/science.ns-4.93.453

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. (2012, January). The Solar Dynamics Observatory (SDO). *Solar Phys.*, *275*(1-2), 3-15. doi: 10.1007/s11207-011-9841-3

Plutino, N., Berrilli, F., Del Moro, D., & Giovannelli, L. (2023). A new catalogue of solar flare events from soft x-ray goes signal in the period 1986–2020. *Advances in Space Research*, *71*(4), 2048-2058. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0273117722010420` (Recent progress in the physics of the Sun and heliosphere) doi: https://doi.org/10.1016/j.asr.2022.11.020

Plutino, N., Michele, B., Grégoire, F., Berrilli, F., Del Moro, D., Giovanelli, L., ... Barata, T. (2024, January). *Solar flare catalog - plutino extension.* Zenodo. Retrieved from

https://doi.org/10.5281/zenodo.10560189 doi: 10.5281/zenodo.10560189
Schrijver, C. J. (2009, March). Driving major solar flares and eruptions: A review. *Advances in Space Research*, *43*(5), 739-755. doi: 10.1016/j.asr.2008.11.004
Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016, October). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *arXiv e-prints*, arXiv:1610.02391. doi: 10.48550/arXiv.1610.02391
Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014, December). Striving for Simplicity: The All Convolutional Net. *arXiv e-prints*, arXiv:1412.6806. doi: 10.48550/arXiv.1412.6806
Sun, Z., Bobra2, M. G., Wang, X., Wang, Y., Sun, H., Gombosi, T., ... Hero, A. (2022, June). Predicting Solar Flares Using CNN and LSTM on Two Solar Cycles of Active Region Data. *Astrophys. J.*, *931*(2), 23. doi: 10.3847/1538-4357/ac64a6
Tan, M., & Le, Q. V. (2021). *Efficientnetv2: Smaller models and faster training.*
van der Sande, K., Flyer, N., Berger, T. E., & Gagnon, R. (2022, November). Solar flare catalog based on SDO/AIA EUV images: Composition and correlation with GOES/XRS X-ray flare magnitudes. *Frontiers in Astronomy and Space Sciences*, *9*, 354. doi: 10.3389/fspas.2022.1031211
Yi, K., Moon, Y.-J., Lim, D., Park, E., & Lee, H. (2021, March). Visual Explanation of a Deep Learning Solar Flare Forecast Model and Its Relationship to Physical Parameters. *Astrophys. J.*, *910*(1), 8. doi: 10.3847/1538-4357/abdebe
Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015, December). Learning Deep Features for Discriminative Localization. *arXiv e-prints*, arXiv:1512.04150. doi: 10.48550/arXiv.1512.04150