

# Multimodal Flare Forecasting with Deep Learning

G. Francisco<sup>1,2,3</sup>, S. Guastavino<sup>4</sup>, J. Fernandes<sup>5</sup>, T. Barata<sup>2</sup> and D. Del Moro<sup>1</sup>

<sup>1</sup> Department of Physics, University of Rome Tor Vergata, Rome, Italy  
e-mail: gregoire.francisco@gmail.com

<sup>2</sup> IA, Instituto De Astrofísica E Ciências Do Espaço, University of Coimbra, Coimbra, Portugal

<sup>3</sup> Department of Physics, University of Rome La Sapienza, Rome, Italy

<sup>4</sup> MIDA, Dipartimento di Matematica, Università di Genova, Genova, Italy

<sup>5</sup> CITEUC, Geophysical and Astronomical Observatory, University of Coimbra and Department of Mathematics, Coimbra, Portugal

October 22, 2024

## ABSTRACT

Solar flare forecasting mainly relies on photospheric magnetograms and associated physical features to predict forthcoming flares. However, it is believed that flare initiation mechanisms often originate in the chromosphere and the lower corona. In this study, we employ deep learning as a purely data-driven approach to compare the predictive capabilities of chromospheric and coronal UV and EUV emissions across different wavelengths with those of photospheric line-of-sight magnetograms. Our findings indicate that individual EUV wavelengths can provide discriminatory power comparable or better to that of line-of-sight magnetograms. Moreover, we identify simple multimodal neural network architectures that consistently outperform single-input models, showing complementarity between the flare precursors that can be extracted from the distinct layers of the solar atmosphere. To mitigate potential biases from known misattributions in Active Region flare catalogs, our models are trained and evaluated using full-disk images and a comprehensive flare event catalog at the full-disk level. We introduce a deep-learning architecture suited for extracting temporal features from full-disk videos.

Use \titlerunning to supply a shorter title and/or \authorrrunning to supply a shorter list of authors.

## 1. Introduction

Solar flares are sudden bursts of electromagnetic radiation and energetic particles that can pose significant threats to human health and technology. Their potential danger is typically assessed based on the intensity of the emitted soft X-Rays (SXR) flux. Flares are classified into five categories: A, B, C, M, and X, with each class representing a peak SXR flux that is one order of magnitude higher than the previous class. The M-Class threshold marks the point from which flares are considered strong, while the X-Class threshold indicates flares that present the most serious threats to society. A significant research effort is given on identifying physical precursors of flares, often derived from photospheric magnetograms. For example, the FLARECAST project identified 209 potential flare precursors, with 94% of them related to Active Region (AR) magnetic field properties (Georgoulis et al. (2021)) Another important trend, initiated by Huang et al. (2018), involves using deep learning as a purely data-driven approach, employing neural networks to magnetogram images. In contrast, the potential of chromospheric and coronal observations for flare prediction remains less explored despite the belief that flare-triggering mechanisms originate in these regions (Shibata & Magara 2011; Toriumi & Wang 2019). One barrier to investigating the upper atmospheric layers for flare precursors is the current observational limitations in deriving magnetograms and related physical quantities from these regions. However, this challenge may soon be addressed at the chromospheric level with upcoming projects like the Solar Activity Monitor Network (SAMNet) Erdélyi et al. (2022)

and the Global Automatic Telescopes for Exploring the Sun (GATES) Giovannelli et al. (2020); Konow et al. (2024) which aim to provide high temporal coverage of chromospheric magnetograms. Moreover, instruments like the Solar Dynamic Observatory (SDO) (Pesnell et al. (2012)) and its Atmospheric Imaging Assembly (AIA) (Lemen et al. (2012)) already provide rich spectral data that may offer valuable thermal and morphological plasma features for both the chromosphere and the corona. In this context, Dissauer et al. (2023) introduced a dataset of AR AIA patches, from which Leka et al. (2023) derived chromospheric and coronal features based on moment analysis, achieving promising results. Notable precursors identified include total emission, steep brightness variations, and high-order moments of running differences, which suggest a tendency for short, small-scale brightening events in flare-imminent regions. Additionally, Sun et al. (2023) utilized Convolutional Neural Network (CNN)s with AIA coronal images to forecast flares above the M-class threshold, achieving state-of-the-art results. Both Leka et al. (2023) and Sun et al. (2023) identified the 94Å line emission as particularly predictive of upcoming flares. Sun et al. (2023) further demonstrated that averaging predictions of single-wavelength models outperforms individual models' predictions. The performance improvement achieved by ensembling models, such as model bagging, can typically be attributed to the law of large numbers. Specifically, under the assumption of independent and identically distributed Gaussian errors, averaging models' outputs reduces error variance, as individual errors tend to cancel each other out. However, this method does not exploit

arXiv:2410.16116v1 [astro-ph.SR] 21 Oct 2024

potential dependencies between the distinct models' features, which might also enhance predictions.

In this study, we aim to extend the works of Leka et al. (2023) and Sun et al. (2023) by:

1. Comparing the discriminative power of chromospheric and coronal features with those derived from line-of-sight magnetograms.
2. Investigating potential linear and non-linear complementarities between features from different atmospheric layers using multimodal models.
3. Exploring the discriminative power of long-term temporal features in addition to single-timestep features.

We utilize Deep Learning as a purely data-driven approach on full-disk images produced by the SDO. This full-disk approach complements the AR-level approach of Leka et al. (2023) and Sun et al. (2023), and reduces mislabels due to known misattributions in AR datasets. Such misattributions, as highlighted by van der Sande et al. (2022), can affect up to 8% of flares above the M-class threshold, and up to 20% of these events are simply missing in the standard GOES Flare catalog. Although the Heliophysics Event Knowledgebase (HEK) corrects many of these issues, some flares remain misattributed or unaccounted for, potentially compromising model training and evaluation in the already challenging context of imbalanced and scarce data. Our focus is on binary forecasts of the occurrence of at least one flare above the M-class threshold within 24-hour windows. This paper is organized as follows: Section 2 describes the data, Section 3 describes the models and training methods, Section 4 describes the results and discussions.

## 2. Data

We use the SDO-2H-ML image dataset introduced in Francisco et al. (2024). This is a 2 hour resolution dataset ranging from May 2010 to April 2023, for a total amount of about 54000 dates for which are associated the corresponding Line-Of-Sight (LOS)-magnetogram and AIA images in the wavelength 1600Å, 304Å, 171Å, 193Å, 211Å and 94Å. The AIA images are pre-processed through alignment, exposure normalization, correction for instrument degradation, and are then downsampled to a resolution of 1024x1024 pixels. AIA and HMI pixel values undergo logarithmic scaling and are reduced to 8-bit depth, preserving the majority of the original pixel distribution with minimal information loss, while also reducing the dataset's size. The image are then compressed as JPEGs, resulting in another moderate loss of small-scale and high frequency details. In an ablation study, we found in Francisco et al. (2024), that such JPEG compression, as well as a further downsampling to 448x448 pixels, had no significant impact on the resulting model performances. We therefore work again with the 448x448 resolution. We use the same flare catalog as in Francisco et al. (2024), which is an extension of Plutino et al. (2023)'s catalog.

## 3. Methodology

### 3.1. Models

#### 3.1.1. Single Timestep Models

To predict flares using single-wavelength images or magnetograms, we employ transfer learning from models pre-trained on the ImageNet dataset Deng et al. (2009). The extensive sample and class diversity in ImageNet enable models to

learn general hierarchical filters and features that transfer well to different problems and datasets. Research by Kornblith et al. (2019) indicates that for new problems with limited data, fine-tuning all layers of a pre-trained model on the new dataset yields optimal performance, typically proportional to the model's original performance on ImageNet. Consequently, we use the EfficientNetV2-S Tan & Le (2021). As of now, EfficientNetV2 models are among the best-performing CNN architectures, significantly outperforming ResNet models He et al. (2015). EfficientNetV2 incorporates advanced residual blocks, including Mobile Inverted Bottleneck Convolution (MBConv) Sandler et al. (2018), which captures complex features with fewer parameters by expanding and then reducing the feature number. Additionally, Fused-MBConv blocks, introduced for that model, optimizes MBConv by merging pointwise and depthwise convolutions into a single step. These innovations contribute to the model's parameter efficiency and performance. The EfficientNetV2 also appears more adapted to our problem than state-of-the-art Vision Transformers (ViTs) Dosovitskiy et al. (2020), such as Swin Liu et al. (2021b) and SwinV2 Liu et al. (2021a), as they only provide slightly better performance, while being harder to fine-tune with small datasets. Indeed, ViTs generally require larger datasets, extensive data augmentation, and regularization techniques Steiner et al. (2021), which can also reduce the effectiveness of transfer learning Kornblith et al. (2019). We choose the EfficientNetV2-S variant with 20 million parameters as it offers a good compromise between performance and ease of fine-tuning for our relatively small training dataset. We remove the top convolutional and prediction layers and replace them with a final convolutional layer with 16 filters, followed by batch normalization and Swish activation. This configuration is more suited to our binary classification problem and smaller dataset than the original 1280 filters. This final convolution block is followed by a global pooling layer, resulting in 16 features that are fed to a final dense layer with 2 neurons and a softmax activation, outputting the probabilities for the negative and positive classes.

To adapt grayscale images to the model, which is pre-trained on RGB images, we duplicate each grayscale image twice to create a 3-channel input. Each resulting model will be denoted by  $Efn-w$ , where  $w$  is the corresponding wavelength in Ångstroms, and  $Efn-B_{LOS}$  for the magnetogram.

For single-timestep multimodal inputs, we compare the following approaches:

1. *Pretrained EfficientNetV2-S on RGB combinations of single inputs.* This approach leverages the transfer learning capabilities of the model originally trained on RGB images. We compare two combinations: a full coronal combination of 193 Å, 211 Å, and 94 Å wavelengths ( $Efn-193x211x94$ ), and a cross-atmospheric combination of the magnetogram, 304 Å, and 94 Å wavelengths ( $Efn-B_{LOS}x304x94$ ).
2. *Features fusion from single-wavelength models.* In this approach, we combine features extracted from separately trained single-input models into a unified fully connected layer to investigate the linear complementarity between features from different inputs. Specifically, we perform this fusion for the cross-atmospheric combination of the magnetogram, 304 Å, and 94 Å wavelengths. The resulting model is denoted  $EfnFuse-B_{LOS}x304x94$ .

### 3.1.2. Video Models

To incorporate temporal dynamics in analyzing the best-performing inputs, we study models using videos composed of 13 frames spaced 2 hours apart, covering the 24 hours preceding the forecasting window. Previous studies, such as Guastavino et al. (2022); Guastavino et al. (2023), have performed flare forecasts on magnetogram videos of ARs using combinations of 2D-CNN and Long Short-Term Memory (LSTM), while Sun et al. (2022) utilized 3D-Convolutions. Our study, however, focuses on full-disk videos rather than ARs and compares EUV multimodal inputs to magnetograms alone. Full-disk videos present unique challenges, and we propose a novel deep learning model designed to effectively infer local solar events from such videos. While 2D convolutions and associated downsampling techniques—such as strides and max pooling—emphasize frame-dominant features, these features may originate from different ARs across distinct full-disk frames. This can complicate the learning of meaningful temporal patterns when using subsequent timeseries models like LSTMs, as the resulting time series of features may correspond to different ARs at different timesteps. In contrast, 3D convolutions better preserve the temporal coherence of features across frames. We propose a hybrid approach, Video Local Event Neural System (VideoLENS), for full-disk videos that maintains this temporal coherence and captures short-term patterns using 3D convolutions, while also leveraging long-term dependencies through LSTM cells and an attention mechanism (Vaswani et al. (2017)). The VideoLENS architecture is illustrated in Figure 1 and detailed in Table 1. The model begins with an initial block of 3D convolutions (*C3D Block*) that scales the original full-disk input down to features with a spatial scale comparable to ARs. Specifically, the *C3D Block* is designed so that each pixel in the final output feature map has an original spatial receptive field slightly larger than the largest possible ARs, accounting for their rotation during the 24 hours covered by the 13 input frames. As shown in Figure 1, each final time series of features (indicated in dark orange) consequently contains information localized to the corresponding original receptive field (shown in blue), ensuring that the time series accurately represents the temporal evolution of the corresponding solar region. A *Local-Timeseries Block* is then applied to each of these feature timeseries to derive more complex temporal features then used for local predictions. This block starts with a multi-head attention layer that computes attention scores along the temporal dimension, ensuring that each timestep is contextually aware of the entire time series and emphasizing the most relevant parts. An LSTM layer is then used to learn long- and short-term features localized to ARs. Subsequently, a pixel-local prediction layer, consisting of a fully connected neuron with sigmoid activation, predicts the probability of a flare at each spatial region. The final full-disk prediction is derived as the maximum of these local predictions, obtained through a GlobalMaxPooling layer that aggregates the local predictions. The resulting model can therefore also be used as a semi-supervised framework able to learn spatial label while only receiving non spatial one at training. It thus extend the Patch-Distributed-CNN (P-CNN) presented in Francisco et al. (2024) to video input, while removing the limitations of the patches artificial boundaries of this former model. Similarly to the P-CNN, precise position estimation of the predicted events could then be calculated from gradient methods for each of the local prediction.

### 3.2. Training and hyperparameters

We use the same full disk Cross-Validation (CV) method as presented in Francisco et al. (2024). Specifically temporal buffers of 27-days are used to create independent temporal chunks that are selected by optimisation to build balanced training and validation folds. This result in a 5-fold CV done on the period ranging from May 2010 to December 2019, while the period from January 2020 to April 2023 is unaltered, meaning every sample and the natural distribution of the data is preserved and kept as a complementary test set. Our models are trained using the Adam optimizer Kingma & Ba (2014). We conducted a Bayesian parameter search for the EfficientNetV2 model with magnetogram inputs ( $Efn-B_{LOS}$ ) to determine an optimal learning rate and the potential inclusion of a weight decay parameter. Although decoupled weight decay regularization Loshchilov & Hutter (2017) is typically employed to reduce overfitting risk, we found it to have minimal impact on the fine-tuning of the pre-trained EfficientNetV2. This observation is consistent with the findings of Kornblith et al. (2019), which indicated that regularization methods may not always be useful in various transfer learning scenarios. Our results also showed that fine-tuning the EfficientNetV2 generally converges in fewer than 10 epochs across most parameter combinations tested. This rapid convergence aligns with Kornblith et al. (2019), who reported that fine-tuning pre-trained models typically requires significantly fewer epochs than training from scratch—approximately 17 times faster. Based on the parameter search results, we train all EfficientNetV2-based models with a learning rate of  $3 \times 10^{-4}$  for 15 epochs and without weight decay regularisation. For the VideoLENS models, a manual parameter search revealed that a slight weight decay regularization of  $1 \times 10^{-5}$  marginally enhances performance. We also found an optimal learning rate that remains the same as for the EfficientNetV2 models, at  $3 \times 10^{-4}$ . Finally, we use weighted binary cross-entropy as the loss function to fully address class imbalance.

### 3.3. Evaluation

To compare the predictive power of the models, we focus on the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC), the True Skill Statistic (TSS), the Heidke Skill Score (HSS) and the Matthews Correlation Coefficient (MCC). The ROC curve gives a model’s achievable recall as a function of the False Positive Rate (FPR) over the different decision thresholds. It directly relates to the TSS which is one of the preferred metrics in flare forecasting studies (Bloomfield et al. (2012)), as the vertical distance between the diagonal and the ROC correspond to the TSS at the corresponding threshold. Optimal ROC AUC values therefore improve the likelihood of finding a threshold with good TSS values. The ROC AUC is an easily interpretable value of a classifier’s discriminatory power as it corresponds to the probability that among two random positive and a negative samples, the classifier will assign a higher probability output to the positive one. The TSS, (Peirce (1884), Hanssen & Kuipers (1965), Woodcock (1976)) is defined as the difference between the recall and the FPR. Also known as the (book-maker) informedness, Peirce’s index or Youden’s J index, it is also equal to the balanced accuracy rescaled between -1 and 1 :

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} = TPR - FPR, \quad (1)$$

**VideoLENS Diagram**

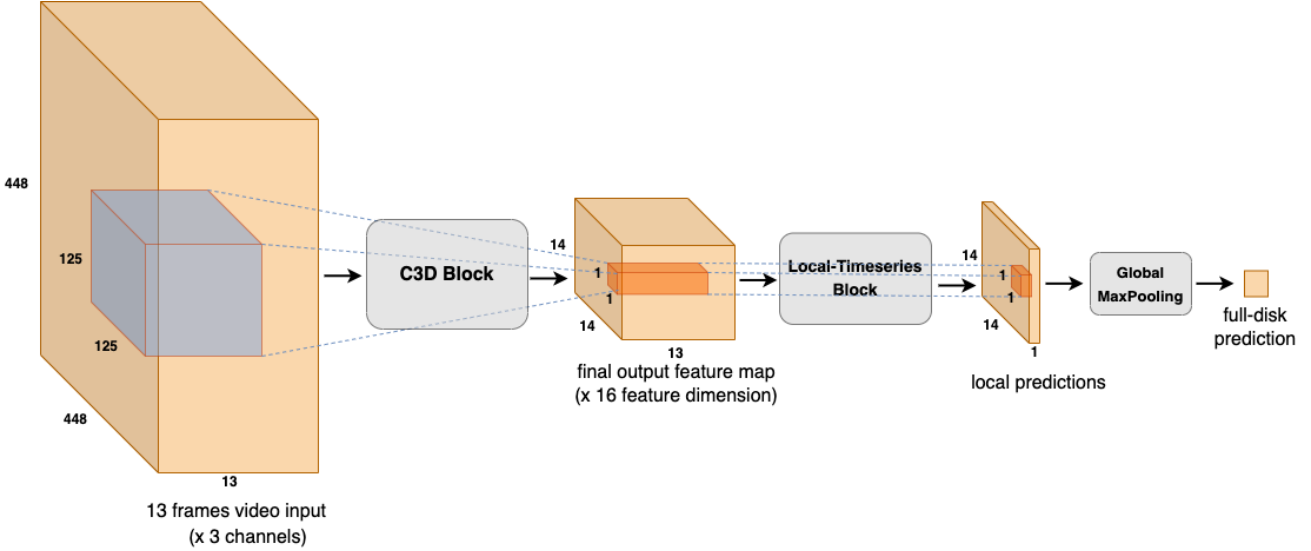


Fig. 1: Video Local Event Neural System (VideoLENS) Diagram. The architectures of the C3D and Local-Timeseries blocks are detailed in Table 1. The blue initial block represents the original receptive field of the time series shown in dark orange. The final local predictions are derived from these features, providing event predictions that are localized to their respective original receptive fields.

Table 1: VideoLENS Architecture

Layer	Kernel	Stride	Regularisation	Activation	Output
<i>C3D block</i>					(13, 14, 14, 16)
Convolution3D (1)	16 x (3, 3, 3)	(1, 2, 2)	BN	Swish	(13, 224, 224, 16)
MaxPooling3D	(3, 3, 3)	(1, 2, 2)	$\emptyset$	$\emptyset$	(13, 112, 112, 16)
Convolution3D (2)	16 x (3, 3, 3)	(1, 2, 2)	BN	Swish	(13, 56, 56, 16)
Convolution3D (3)	16 x (3, 3, 3)	(1, 1, 1)	BN	Swish	(13, 56, 56, 16)
Convolution3D (4)	8 x (3, 3, 3)	(1, 1, 1)	BN	Swish	(13, 56, 56, 8)
Convolution3D (5)	8 x (3, 3, 3)	(1, 1, 1)	BN	Swish	(13, 56, 56, 8)
Convolution3D (6)	8 x (3, 3, 3)	(1, 2, 2)	BN	Swish	(13, 28, 28, 8)
Convolution3D (7)	8 x (3, 3, 3)	(1, 1, 1)	BN	Swish	(13, 28, 28, 8)
Convolution3D (8)	8 x (3, 3, 3)	(1, 1, 1)	BN	Swish	(13, 28, 28, 8)
Convolution3D (9)	8 x (3, 3, 3)	(1, 1, 1)	BN	Swish	(13, 28, 28, 8)
Convolution3D (10)	16 x (3, 3, 3)	(1, 2, 2)	BN + DP(0.2)	Swish	(13, 14, 14, 16)
<i>Local Timeseries Block</i>					(14, 14, 1)
MultiHead-Attention	4-head x 16-dim	residual connection	LN + DP(0.2)	$\emptyset$	(13, 16)
LSTM	16-cells	$\emptyset$	BN	$\emptyset$	(16)
Dense (local predication)	1	$\emptyset$	$\emptyset$	Sigmoid	(1)
GlobalMaxPooling	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$	(1)

Architecture of the VideoLENS. BN : Batch Normalisation. LN : Layer Normalisation. DP(0.2) : dropout with a rate of 0.2. The *Local Timeseries Block* is applied to each spatial-point of the attention layer output, i.e. to every  $(13, i, j, 16)$  with  $(i, j) \in \llbracket 1, 14 \rrbracket^2$ .

or equivalently

$$TSS = TPR + TNR - 1, \quad (2)$$

where TPR and TNR represent the true positive rate and true negative rate, respectively. It is noted that the TSS and ROC AUC

suffer some limitations in measuring a model's ability to discriminate properly between two classes.



The two metrics only encompass information about the success rate of detection in each class, without information about the precision with which predictions are made. In imbalanced cases, this flaw allows high TSS and ROC AUC to mask models that distinguish between negative and positive with poor precisions (Jeni et al. (2013)). In the specific case of flare forecasting, some studies, exhibit TSS and ROC AUC that are higher for M+ forecasting than C+ forecasting on the same datasets (e.g. Guastavino et al. (2022), Leka et al. (2023)). This gives the misleading impression that the models discriminate better stronger flare than medium ones, whereas they usually have precision at least twice as small when predicting these stronger events. The HSS, defined in equation 3, was originally introduced by Heideke (1926) for evaluating weather forecasts. Also known as Cohen’s Kappa index, it is widely used in flare forecasting to compare a model’s skill relative to a random guess model (Camporeale (2019)). The HSS is particularly comprehensive for imbalanced datasets, as it combines information from the TSS with the Markedness, which synthesizes the class-wise precision of the model’s predictions for each class, including both Positive Predictive Value (PPV) and Negative Predictive Value (NPV). The Markdness is the precision analog to how the TSS captures class-wise accuracy, synthesizing True Positive Rate (TPR) and True Negative Rate (TNR). For a deeper understanding of how the HSS integrates both TSS and Markdness, one can refer to the mathematical analysis presented in Delgado & Tibau (2019). This analysis highlights the close relationship between the HSS and the binary correlation coefficient, which serves as a geometric average between TSS and Markdness. Specifically, from the  $\alpha$  and  $\beta$  coefficients in Delgado & Tibau (2019), the HSS can be understood as the harmonic mean of Informedness (TSS) and Markedness, adjusted for the model’s frequency bias.

$$HSS = 2 * \frac{TP * TN - FN * FP}{P(TN + FN) + N(TP + FP)}, \quad (3)$$

where  $P$  and  $N$  represents the total number of positive and negative samples, respectively. The MCC (Matthews (1975)), defined in equation 4, is the Pearson correlation coefficient between two binary variables. It contains similar information to the HSS with constant even weighting between TSS and the Markdness. It thus provides an agnostic measure of a model discriminatory power as it takes into account all the basic confusion matrix rate of a model giving equal importance to each of them. This makes it particularly suitable to assess models discriminatory powers on imbalanced problems. We provide a more detail comparison of the TSS, HSS and MCC and discussion of their strength and limitations in Francisco et al. (2024) (section 2.1.1 of and associated appendix).

$$MCC = \frac{TP * TN - FN * FP}{\sqrt{P(TN + FN) * N * (TP + FP)}} \quad (4)$$

## 4. Results & Discussions

Figures 2, 4, 3 and 5 present the performance metrics of the models: ROC AUC, HSS, TSS and MCC respectively. The plots rank the models based on their performance, with a consistent ranking observed across all three metrics. Notably, while ROC AUC values are clustered within a narrow range (+16% from the worst to the best model), the MCC, HSS and TSS demonstrate more pronounced disparities in model performance (with maximum performance gaps of +54% and +51%, respectively). The results

also show stability between validation and test sets, indicating good generalization capabilities. However, the model using the 1600Å wavelength exhibited a marked decline in HSS and TSS by 24% and 26%, respectively, on the test set compared to the validation set. This performance drop may be attributed to the learning of irrelevant features from the 1600Å wavelength that persist across the 27-day buffer period of the CV folds, leading to potential overfitting on the validation data. In addition, there might be some pattern specific to the solar cycle’s 24th period which does not generalize to the next cycle. This finding underscore that even rigorous CV methods do not guarantee performance generalization in operational settings. Consequently, incorporating a separate operational test set, as advocated by Cinto et al. (2020), appears essential for a more comprehensive evaluation of model performance.

### 4.1. Coronal and Chromospheric Features are Most Predictive of Flares

The 94Å wavelength provides the most discriminative features for flare prediction, consistent with findings from Leka et al. (2023) and Sun et al. (2023). Compared to the line-of-sight magnetograms, the *Efn-94* model shows a significant improvement, with a 3.4% increase in TSS on the test set and a 10.4% increase on the validation set. Additionally, HSS values improve by 12% on the test set and 9% on the validation set. Although the 193Å and 211Å wavelengths demonstrate similar or lower predictive performance to the line-of-sight magnetograms, their combination with the 94Å wavelength in the *Efn-193x211x94* model results in a further enhanced HSS, showing an overall increase of 18% on both the test and validation sets compared to the magnetograms alone. This suggests that the relative intensity variations between these three wavelengths may serve as effective precursors for flare events. Chromospheric observations also contribute valuable predictive features. While the lower chromospheric wavelength of 1600Å exhibits weak generalization performance on the test set, the upper chromospheric wavelength of 304Å emerges as the second most predictive single input, following the 94Å wavelength.

### 4.2. Features from Different Atmospheric Layers Complement Each Other

Combining features from various solar atmospheric layers enhances overall performance compared to using the single best-performing layer. The model *EfnFuse-BLOSx304x94*, which integrates features from *Efn - BLOS*, *Efn - 304*, and *Efn - 94*, outperforms the *Efn - 94* model by 5.7% and 7.4% in HSS on the test and validation sets, respectively. While improvements in TSS and ROC AUC are more modest, the notable increase in HSS indicates enhanced precision, with improved performances in discriminating ambiguous cases. We found that, cross-atmospheric feature complementary is most effective through feature fusion after training individual models for each input (*EfnFuse-BLOSx304x94*). Training a single model directly on the cross-atmospheric channel combination with the pretrained EfficientNetV2 (*Efn-BLOSx304x94*) does not surpass the performance of the model using the most predictive channel alone, conversely to the coronal wavelength combination (*Efn-193x211x94*). This result might arise from the 2D-CNN architecture, where combining multiple channels in early convolutional layers treats them as a distinct feature dimension.

## ROC AUC

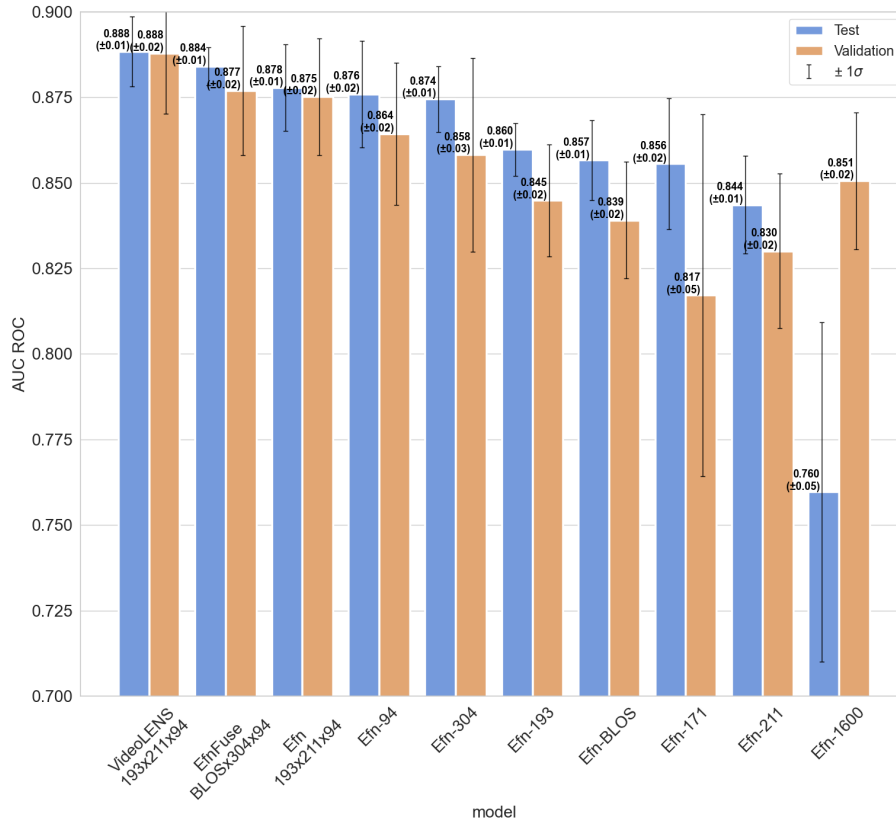


Fig. 2: ROC AUC of various models. Efn-*input* denotes the EfficientNetV2-S model trained on individual frame *input*. EfnFuse-*inputs* refers to a logistic regression model utilizing features extracted by EfficientNetV2-S models trained on each of the distinct *inputs*. VideoLENS represents the video-based models.

## TSS

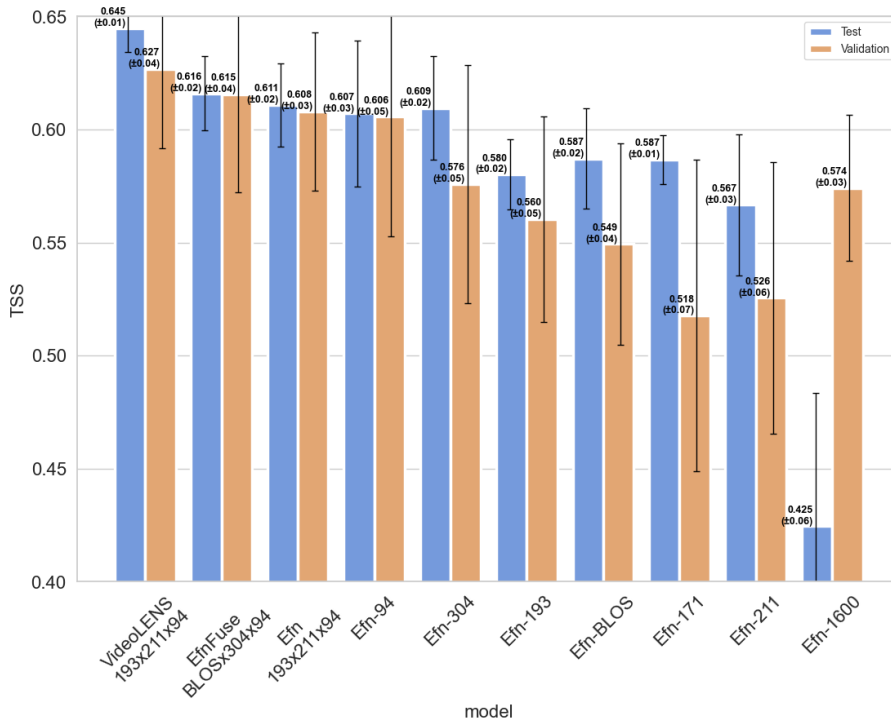


Fig. 3: Models' TSS. Models' labels are the same as the ones described in 2.

### HSS

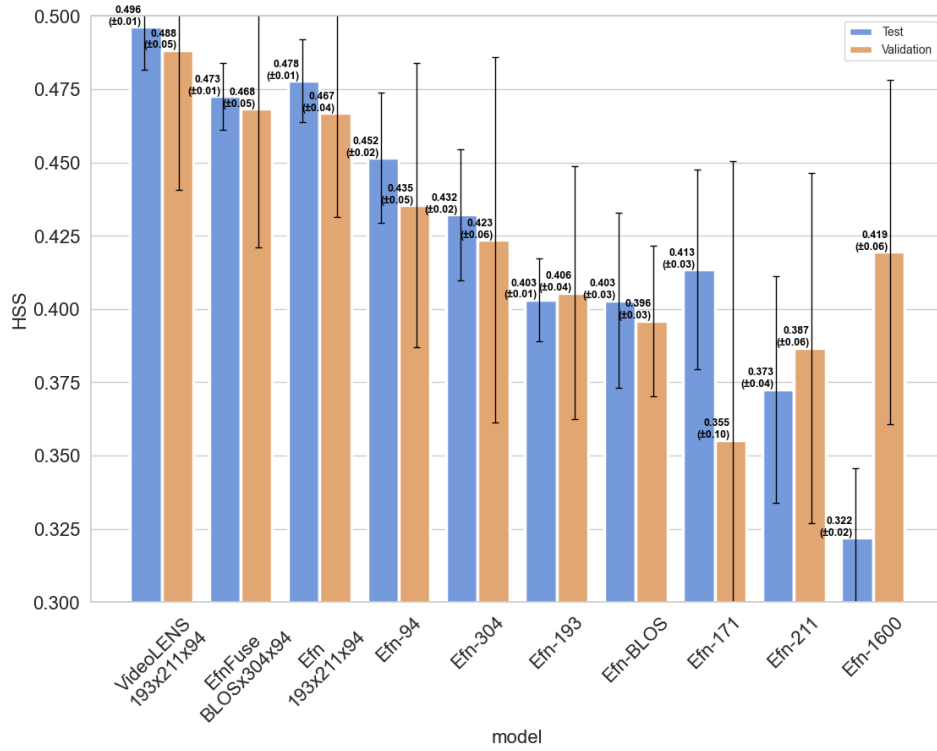


Fig. 4: Models' HSS. Models' labels are the same as the ones described in 2.

### MCC

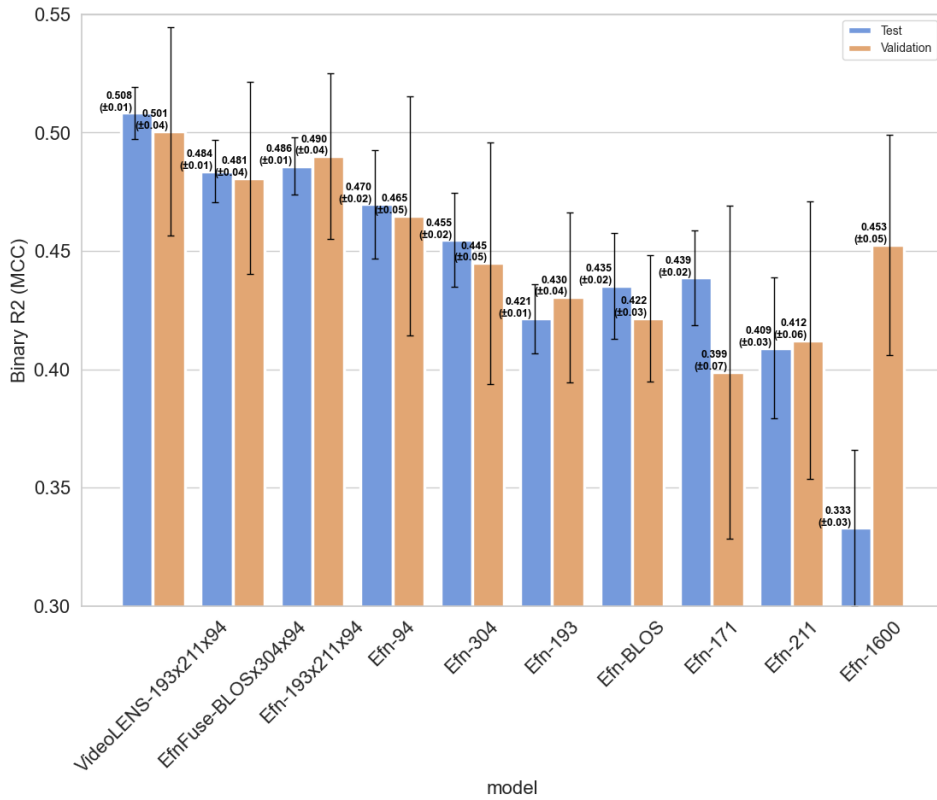


Fig. 5: Models' MCC. Models' labels are the same as the ones described in 2.



Filters learn different weights for each channel, and sum their outputs to form the feature map. This approach is particularly effective for features that complement each other linearly, such as the coronal wavelength channels in the *Efn-193x211x94*, which capture different emission lines of similar coronal structures. In contrast, magnetograms, chromospheric images, and coronal images — representing more distinct physical layers and processes — might not combine linearly as well in the first convolutional layers. Alternative approaches to improve cross-atmospheric models performance could include using 3D convolutions in the initial layer of a single model, or training on all layers simultaneously the distinct block of the *EfnFuse-BLOS x304x94*, rather than pre-training them separately. However, this latter method may require larger datasets or data augmentation to prevent overfitting due to the increased model complexity. Ultimately, while combining features from different atmospheric layers enhances performance, it does not significantly surpass the results achieved with the three coronal channels in the *Efn-193x211x94* model.

#### 4.3. Temporal Dynamics Enhance Predictive Performance Compared to Static Features

The *VideoLENS-193x211x94* model outperforms all others, showing up to a 5.8% improvement in ROC AUC, a 14% increase in TSS, and a 23% enhancement in HSS over the *Efn-BLOS* model. When compared to the static single-frame model *Efn-193x211x94*, the *VideoLENS-193x211x94* achieves improvements of up to 5.6% in TSS and 4.4% in HSS. Attempts to combine other wavelength pairs, such as the cross-atmospheric combination *BLOS x304x94*, using both parallel processing and an RGB-like processing approach similar to *VideoLENS-193x211x94*, did not yield models significantly surpassing the performance of the *Efn-193x211x94* single-frame coronal model. These results underscore that the 193x211x94 wavelength combination provides some of the most predictive information for flare forecasting among the atmospheric layer combinations tested. The *Efn-193x211x94* and *VideoLENS-193x211x94* models achieve HSS scores of 0.46 and 0.48, respectively, significantly surpassing other known full-disk models for forecasting M+ flares within 24 hours. For context, Pandey et al. (2023) reports an HSS of 0.35 for a similar problem using full-disk line-of-sight magnetograms over a comparable validation period.

#### 4.4. VideoLENS Ablation Study

The VideoLENS architecture was evaluated through an ablation study to determine the impact of various components on model performance. This study revealed that incorporating local-wise predictions and the time-series block offered slight improvements in ROC AUC compared to models using only C3D with global pooling. Increase in model size, through the number of filters or parameters of the attention and LSTM layers, did not result in significant performance gains, likely due to the modest size of the training dataset. At equal number of parameters, fewer C3D layers but more filters performed less effectively, indicating that simply increasing the number of filters is not beneficial without careful architectural tuning. More importantly, every C3D-based models consistently outperformed 2D-CNN-LSTM architectures, highlighting that the latter may not be as effective for analyzing full-disk videos. This is attributed to the tendency of the 2D-CNN component to capture features from various ARs

across different frames, when multiple ARs coexist during the video.

#### 4.5. The last and first flare challenge

An examination of model performance revealed that all models exhibited roughly null TSS and HSS scores when evaluated on the samples' subsets where the label of the predicted time window differed from the previous time window. This confirms that our previous observations from Francisco et al. (2024) generalise to a wide range of input modalities and to the use of low temporal resolution data to forecast flares. Specifically, the models struggle to forecast events that occurred in the next 24 hours following a period of inactivity, as well as to forecast a quiet period when flares occurred in the past 24 hours. This indicates that while the models can effectively predict flares when there is a persistent level of activity, they face substantial difficulties in scenarios involving activity changes, with performance akin to random guessing in these transitional cases. The current most predictive flare precursors might, therefore, be more indicative of the current activity level of an AR rather than being actually predictive.

## 5. Conclusion and Future Works

For years, studies on solar activity forecasting have been employing the photospheric AR complexity (computed either from continuum images or from magnetic field information) to predict the AR flare activity (e.g., Schrijver 2007; Korsós et al. 2015; Benvenuto et al. 2018; Korsós et al. 2020; Cicogna et al. 2021). However, it is widely accepted that the coronal magnetic topology and dynamics are crucial to driving and then triggering the release of solar eruptions.

Since direct information on the coronal magnetic field is still not available for such studies, researchers are using as a proxy the EUV multiwavelength images to reveal the onset mechanism of solar flares in case studies (Gosain 2012; Imada et al. 2014; Bamba et al. 2014) and also for statistic approaches (Dissauer et al. 2023; Leka et al. 2023). Such uses have been recently extended to deep-learning approaches to automatically extract flare precursors from EUV images to build models for flare forecasting (e.g., Sun et al. 2023; Nishizuka et al. 2020). Following this trend, we created a suitable dataset for DL approaches and investigated different 2D and 3D CNN models, exploiting the multiwavelength EUV images from the AIA instrument, the photospheric LoS magnetogram from the HMI instrument, and the temporal dimension.

Below is a summary of the main findings from this study:

- 1. Enhanced Forecasting with Coronal Wavelengths:** The use of coronal EUV wavelengths emissions notably enhances forecasting performance compared to relying on photospheric line-of-sight magnetograms. The combination of emission lines at 193 Å, 211 Å, and 94 Å proves particularly effective. This suggests that the relative intensity between EUV wavelengths may provide crucial information about flare precursors. Future research will focus on a detailed investigation of the *Efn-193x211x94* model to identify specific precursors through explainability methods and gain further insights into flare mechanisms.
- 2. VideoLENS Architecture and Temporal Dynamics:** The VideoLENS architecture is an efficient way to forecast localized solar events from full-disk videos. Notably, the use of such model to incorporate temporal dynamics in flare predictive features yields superior performance compared to single-timestep input models. Specifically, the VideoLENS model using the 193 Å, 211 Å, and 94 Å wavelengths achieves an HSS of 0.50 and a TSS of 0.65, substantially outperforming typical full-disk models that forecast M-class flares within the next 24 hours.
- 3. Challenges in Forecasting Activity Changes:** All tested models exhibit limitations in forecasting transitions in activity levels. This suggests that the features derived from the studied inputs are more indicative of current activity levels rather than exclusive predictors of upcoming flares. While the inherent stochasticity of complex physical systems may partly explain these limitations in forecasting activity changes, further research could be of interest.

Future work will explore higher temporal resolution features than the 2 hours used in this work as well as different forecasting window sizes, and the addition of other data modalities such as SXR and Extreme Ultraviolet (EUV) flux timeseries.

*Acknowledgements.* 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 Skłodowska-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. 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

- Bamba, Y., Kusano, K., Imada, S., & Iida, Y. 2014, Publications of the Astronomical Society of Japan, 66, S16
- Benvenuto, F., Piana, M., Campi, C., & Massone, A. M. 2018, The Astrophysical Journal, 853, 90
- Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJ, 747, L41
- Camporeale, E. 2019, Space Weather, 17, 1166
- Cicogna, D., Berrilli, F., Calchetti, D., et al. 2021, The Astrophysical Journal, 915, 38
- Cinto, T., Gradwohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. 2020, MNRAS, 495, 3332
- Delgado, R. & Tibau, X.-A. 2019, PLoS One, 14, e0222916
- Deng, J., Dong, W., Socher, R., et al. 2009, in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255
- Dissauer, K., Leka, K. D., & Wagner, E. L. 2023, ApJ, 942, 83
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2020, arXiv e-prints, arXiv:2010.11929
- Erdélyi, R., Korsós, M. B., Huang, X., et al. 2022, Journal of Space Weather and Space Climate, 12, 2
- Francisco, G., Berretti, M., Chierichini, S., et al. 2024, ESS Open Archive
- Georgoulis, M. K., Bloomfield, D. Shaun, Piana, Michele, et al. 2021, J. Space Weather Space Clim., 11, 39
- Giovannelli, L., Berrilli, F., Calchetti, D., et al. 2020, Journal of Space Weather and Space Climate, 10, 58
- Gosain, S. 2012, The Astrophysical Journal, 749, 85
- Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. 2022, A&A, 662, A105
- Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. 2023, Frontiers in Astronomy and Space Sciences, 9, 1039805
- 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)
- He, K., Zhang, X., Ren, S., & Sun, J. 2015, arXiv e-prints, arXiv:1512.03385
- Heidke, P. 1926, Geografiska Annaler, 8, 301
- Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
- Imada, S., Bamba, Y., & Kusano, K. 2014, Publications of the Astronomical Society of Japan, 66, S17
- Jeni, L. A., Cohn, J. F., & De La Torre, F. 2013, in 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 245–251
- Kingma, D. P. & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980
- Konow, F., Berrilli, F., Calchetti, D., et al. 2024, in Ground-based and Airborne Instrumentation for Astronomy X, Vol. 13096, SPIE, 3007–3016
- Kornblith, S., Shlens, J., & Le, Q. V. 2019, in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (Los Alamitos, CA, USA: IEEE Computer Society), 2656–2666
- Korsós, M., Georgoulis, M., Gyenge, N., et al. 2020, The Astrophysical Journal, 896, 119
- Korsós, M. B., Ludmány, A., Erdélyi, R., & Baranyi, T. 2015, The Astrophysical Journal Letters, 802, L21
- Leka, K. D., Dissauer, K., Barnes, G., & Wagner, E. L. 2023, ApJ, 942, 84
- Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, Sol. Phys., 275, 17
- Liu, Z., Hu, H., Lin, Y., et al. 2021a, arXiv e-prints, arXiv:2111.09883
- Liu, Z., Lin, Y., Cao, Y., et al. 2021b, arXiv e-prints, arXiv:2103.14030
- Loshchilov, I. & Hutter, F. 2017, arXiv e-prints, arXiv:1711.05101
- Matthews, B. 1975, Biochimica et Biophysica Acta (BBA) - Protein Structure, 405, 442
- Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., & Ishii, M. 2020, ApJ, 899, 150
- Pandey, C., Angryk, R. A., & Aydin, B. 2023, in Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track, ed. G. De Francisci Morales, C. Perlich, N. Ruchansky, N. Kourtellis, E. Baralis, & F. Bonchi (Cham: Springer Nature Switzerland), 72–89
- Peirce, C. S. 1884, Science, 4, 453
- Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, Sol. Phys., 275, 3
- Plutino, N., Berrilli, F., Del Moro, D., & Giovannelli, L. 2023, Advances in Space Research, 71, 2048, recent progress in the physics of the Sun and heliosphere
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. 2018, arXiv e-prints, arXiv:1801.04381
- Schrijver, C. J. 2007, The Astrophysical Journal, 655, L117
- Shibata, K. & Magara, T. 2011, Living Reviews in Solar Physics, 8, 1
- Steiner, A., Kolesnikov, A., Zhai, X., et al. 2021, arXiv e-prints, arXiv:2106.10270
- Sun, D., Huang, X., Zhao, Z., & Xu, L. 2023, ApJS, 266, 8
- Sun, P., Dai, W., Ding, W., et al. 2022, ApJ, 941, 1
- Tan, M. & Le, Q. V. 2021, EfficientNetV2: Smaller Models and Faster Training
- Toriumi, S. & Wang, H. 2019, Living Reviews in Solar Physics, 16, 3
- van der Sande, K., Flyer, N., Berger, T. E., & Gagnon, R. 2022, Frontiers in Astronomy and Space Sciences, 9, 354
- Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, arXiv e-prints, arXiv:1706.03762
- Woodcock, F. 1976, Monthly Weather Review, 104, 1209

## 6. List of Acronyms

<b>AR</b>	Active Region
<b>CNN</b>	Convolutional Neural Network
<b>EUV</b>	Extreme Ultraviolet
<b>LOS</b>	Line-Of-Sight
<b>LSTM</b>	Long Short-Term Memory
<b>SDO</b>	Solar Dynamic Observatory
<b>TSS</b>	True Skill Statistic
<b>HSS</b>	Heidke Skill Score
<b>MCC</b>	Matthews Correlation Coefficient
<b>P-CNN</b>	Patch-Distributed-CNN
<b>AIA</b>	Atmospheric Imaging Assembly
<b>SXR</b>	Soft X-Rays
<b>CV</b>	Cross-Validation
<b>VideoLENS</b>	Video Local Event Neural System
<b>AUC</b>	Area Under the Curve
<b>ROC</b>	Receiver Operating Characteristic
<b>FPR</b>	False Positive Rate