A&A 662, A105 (2022)
https://doi.org/10.1051/0004-6361/202243617
© S. Guastavino et al. 2022

Astronomy & Astrophysics

Implementation paradigm for supervised flare forecasting studies: A deep learning application with video data

Sabrina Guastavino[1], Francesco Marchetti[2], Federico Benvenuto[1], Cristina Campi[1], and Michele Piana[1,3]

[1] MIDA, Dipartimento di Matematica, Università degli Studi di Genova, Via Dodecaneso 35, 16146 Genova, Italy
e-mail: guastavino@dima.unige.it
[2] Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova, Padova, Italy
e-mail: francesco.marchetti@unipd.it
[3] CNR-SPIN, Via Dodecaneso 33, 16146 Genova, Italy
e-mail: piana@dima.unige.it

Received 22 March 2022 / Accepted 16 May 2022

ABSTRACT

Aims. In this study, we introduce a general paradigm for generating independent and well-balanced training, validation, and test sets for use in supervised machine and deep learning flare forecasting, to determine the extent to which video-based deep learning can predict solar flares.

Methods. We use this implementation paradigm in the case of a deep neural network, which takes videos of magnetograms recorded by the Helioseismic and Magnetic Imager onboard the Solar Dynamics Observatory (SDO/HMI) as input.

Results. The way the training and validation sets are prepared for network optimization has a significant impact on the prediction performances. Furthermore, deep learning is able to realize flare video classification with prediction performances that are in line with those obtained by machine learning approaches that require an a priori extraction of features from the HMI magnetograms.

Conclusions. To our knowledge, this is the first time that the solar flare forecasting problem is addressed by means of a deep neural network for video classification, which does not require any a priori extraction of features from the HMI magnetograms.

Key words. Sun: flares – Sun: activity – Sun: magnetic fields – methods: data analysis

1. Introduction

Solar flares are the most explosive events in the larger field of space weather (Tandberg-Hanssen & Emslie 1988; Hudson 2011; Shibata & Magara 2011). They are related to a sudden release of magnetic energy, and there is much observational evidence that such a release is a consequence of the reconnection and reconfiguration of magnetic field lines high in the solar corona (Shibata 1996; Sui et al. 2004; Su et al. 2013). However, there is no current agreement on the physical mechanisms that allow the conversion of magnetic energy into nonthermal accelerated particles, nor more generally on the physical model that better explains the whole stream from magnetic reconnection, through flaring emission, down to other space weather manifestations such as coronal mass ejections (CMEs) and solar energetic particles (SEPs; Aschwanden 2008).

In this context, the issue of forecasting solar flares plays a crucial role since it involves open problems in plasma and high-energy physics, modeling of complex systems, and computational science. Therefore, from an astrophysical viewpoint, solving the flare prediction problem would help to identify the basic mechanisms that trigger flares and connect them to space weather (Schwenn 2006; McAteer et al. 2010). Furthermore, from an operational viewpoint, it would lead to a notable improvement in space weather monitoring and the protection of in-space and on-Earth technologies from space weather hazards (Crown 2012; Murray et al. 2017).

The main strategies to address solar flare forecasting involve deterministic models (Strugarek & Charbonneau 2014; Petrakou 2018), statistical methods (Song et al. 2009; Mason & Hoeksema 2010; Bloomfield et al. 2012; Barnes et al. 2016a), and machine learning (see Georgoulis et al. 2021a and references therein). A first comparison of algorithms is contained in a series of four papers that focus particularly on operational systems (Barnes et al. 2016b; Leka et al. 2019a,b; Park et al. 2020). However, in recent years, the impressive development of artificial intelligence has inspired a correspondingly high number of machine and deep learning studies based on the analysis of historical archives of magnetograms. On the one hand, these approaches are providing promising performances, even in operational frameworks (Nishizuka et al. 2018, 2020, 2021). On the other hand, they are opening up several issues, mainly related to their supervised nature and to their applicability to more general space weather contexts (Ahmadzadeh et al. 2021; Georgoulis et al. 2021a,b).

Approaches formulated within the machine learning framework typically need three ingredients: a supervised algorithm for classification, a historical data set for its training, and a score for the assessment of performance. However, one of the most intriguing aspects of the flare forecasting game addressed by machine learning is that the studies performed so far have led to significantly different skill scores, even though they were applied to the same data archive. Just as a very partial example, Table 1 contains the performance outcomes of twelve flare forecasting studies realized by means of machine learning approaches as applied to observations provided by the same space instrument, the Helioseismic and Magnetic Imager onboard the Solar Dynamics Observatory (SDO/HMI). For each of the twelve

Table 1. Description of twelve flare forecasting studies based on machine learning.

Paper	Data	Multiple test realizations	AR data split	Validation	Method	Score – C1+	Score – M1+
Bobra & Couvidat (2015)	Point in time features (SHARP)	Yes	Yes	Yes	SVM	–	0.74
Liu et al. (2017)	Point in time features (SHARP)	Yes	Yes	Yes	RF	–	0.76
Nishizuka et al. (2018)	Point in time features (ad hoc computed)	No	No	No	MLP	0.63	0.80
Florios et al. (2018)	Point in time features (FLARECAST)	Yes	Yes	Yes	RF	0.60	0.74
Jonas et al. (2018)	Time series features	Yes	No	Yes	RF	–	0.74 – 0.81
Campi et al. (2019)	Point in time features (FLARECAST)	Yes	No	Yes	Hybrid lasso	0.54	0.67
Liu et al. (2019)	Time series features (SHARP)	Yes	No	Yes	LSTM	0.61	0.79
Wang et al. (2020)	Time series features (SHARP)	…	No	yes	LSTM	0.55	0.68
Park et al. (2018)	HMI and MDI magnetograms	No	No	Yes	CNN	0.63	…
Huang et al. (2018)	HMI and MDI magnetograms	Yes	Yes	–	CNN	0.49	0.66
Li et al. (2020)	HMI magnetograms	Yes	No	No	CNN	0.68	0.75
Yi et al. (2021)	HMI magnetograms	No	No	Yes	CNN	0.65	–

Notes. For each study, the table reports: the main author (column “paper”); the kind of data used (column “data”); whether a confidence strip has been computed for the skill score (column “multiple test realizations”); whether data belonging to the same AR are split between the training and test sets (column “AR data split”); whether a validation set has been used to optimize the machine learning algorithm (column “validation”); which method has been used (column “method”); and the score values for the prediction of C1+ and M1+ flares (columns “score – C1+” and “score – M1+”).

studies, the table reports: whether a confidence interval on the skill score is computed; whether data belonging to the same active region (AR) are split between the training and test set; whether a validation set is exploited to optimize the machine learning algorithms; which is the machine learning algorithm applied; which are the values obtained for a specific skill score in the case of the prediction of flares with a Geostationary Operational Environmental Satellite (GOES) class higher than C (C1+ flares) and M (M1+ flares).

In our opinion, the reason for the heterogeneity of the skill score values in the table is due to the fact that there is no general agreement among flarecasters on a validation strategy for the prediction methods. Indeed, given a specific historical archive, these twelve methods generate training, validation, and test sets according to completely different rules. Furthermore, not all methods compute a confidence interval in order to assess the statistical reliability of the scores, and not all methods distinguish between validation and testing.

The first objective of the present paper is to propose a general paradigm for the implementation and assessment of flare forecasting processes, based on supervised machine and deep learning approaches. Specifically, we believe that this kind of strategy should suggest a common perspective on two specific issues: the way data preparation is realized, with a specific focus on the way the historical data set is split into a training set, a validation set, and a test set; and the way the prediction performances are presented, with a specific focus on the computation of the statistical reliability of results.

As far as the first issue is concerned, we propose a standardized approach to data splitting that is not based on a chronological criterion. Indeed, chronological splitting introduces a bias among training, validation, and test sets, which is a consequence of the cyclicity of the solar activity. This bias cannot be easily removed since databases at our disposal are limited in time and are typically included in one solar cycle, while it is well established in machine learning theory (Vapnik 1998) that training and test sets should be drawn from the same distribution (this is not the case when chronological splitting is adopted). Therefore, to assess the performances of machine learning methods, we formulate a best practise, relying on machine learning theory. This is done via definition of a data sample that accounts for the uniformity of training, validation, and test sets with respect to both flare classes and the different possible typologies of null events. As far as the second issue is concerned, we point out that any machine learning algorithm should be repeatedly trained on data subsets generated by means of random extractions of ARs from the HMI archive, in order to associate a confidence interval with the skill score values computed on the test set.

The second objective of this paper is to present, for the first time, a deep learning technique that takes HMI videos as input and provides a binary prediction with no intermediate processing of the computed features as output. Our technique is based on a long-term recurrent convolutional network (LRCN) architecture (Yu et al. 2019), which combines the use of a convolutional neural network (CNN), for the extraction of morphological features of the ARs, with a long short-term memory (LSTM) network (Hochreiter & Schmidhuber 1997), for the temporal analysis of the sequences. CNNs for HMI videos have already been used by Chen et al. (2019), following an approach that first extracts features from the magnetograms by means of an autoencoder network, then artificially removes redundant features extracted by the CNN according to a p-value analysis, and finally organizes the extracted features into time series given as input to an LSTM network, which computes a binary prediction.

Unlike the technique used in Chen et al. (2019), our proposed method does not separate the CNN analysis from the LSTM one, and therefore it does not need an a posteriori processing of the features extracted by the CNN. This is a crucial point, since the weights updating process for the autoencoder network in Chen et al. (2019) depends on the optimization of a regression loss, which measures the discrepancy between the reconstructed images and the experimental ones, whereas the weights updating process for the CNN in the LRCN network depends on the optimization of a classification loss, which measures the discrepancy between the predicted probability that an event occurs with the actual YES-NO labels.

This paper is organized as follows: Sect. 2 describes the implementation paradigm; Sect. 3 focuses on such a strategy when applied to video data preparation; Sect. 4 discusses the design of the applied deep learning model; and Sect. 5 is devoted to the description of the prediction results. Our conclusions are offered in Sect. 6.

2. Implementation and assessment paradigm

2.1. Sample definition

Given an AR, we split the data associated with it into contiguous samples, each one corresponding to a time interval of fixed duration. When the interval is reduced to only one time point, each sample can be a set of numerical values, for example, values of physical features of that AR, or an AR image. Alternatively, when the time interval is bigger than one time point, a sample can be a time series of features, or a video of magnetograms. In this study, data samples were labeled as type X, M, or C, depending on the ability of the AR to which the sample belonged to generate a flare of a certain intensity.

- X class sample: a sample of the AR that originated a flare in the 24 h after the sample time, with a maximum flare class of X1 or above;
- M class sample: a sample of the AR that originated a flare in the 24 h after the sample time, with a maximum flare class of M1 or above, but lower than class X;
- C class sample: a sample of the AR that originated a flare in the 24 h after the sample time, with maximum flare class C1 or above, but lower than class M.
  Furthermore, we decided to account for the intrinsic heterogeneity of the null event class by labeling the data samples that corresponded to null events according to four different classes:
- NO1 class sample: a sample of the AR that never originated a C1+ flare;
- NO2 class sample: a sample of the AR that did not originate a C1+ flare in the 24 h after the sample time, that had not originated a C1+ flare in the past, but did originate a C1+ flare in the future;
- NO3 class sample: a sample of the AR that did not originate a C1+ flare in the 24 h after the sample time, but did originate a C1+ flare in the 48 h before the sample time. This is a natural choice, since in this context many relevant features refer to the past 24-h window (e.g., flare index past and flare past);
- NO4 class sample: a sample of the AR that did not originate a C1+ flare in the 24 h after the sample time and did not originate a C1+ flare in the 48 h before the sample time, but did originate a C1+ flare before the 48 h before the sample time.

We can therefore think of an AR as a set of data samples labeled according to the abovementioned criteria. For example, if we suppose that the AR number 12645 includes 4 X samples, 10 M samples, 24 C samples, and 13 NO2 samples, then this AR can be described by means of the notation: $\mathrm{AR}_{12645} = \{4\mathrm{X}, 10\mathrm{M}, 24\mathrm{C}, 13\mathrm{NO2}\}$.

2.2. Well-balanced data sets

The procedure for the generation of training, validation, and test sets was based on two criteria.

- Proportionality: We required the sets to have almost equal rates of samples for each sample type described above. In order to construct the sets as reliably as possible, we required the rates to be similar to those characterizing the historical archive. In our experiments, we therefore set the following rates, which are coherent to those in the HMI archive for the time interval between 2012 September 14 and 2017 September 30 (where p_X denotes the rate of the X class sample, p_M that of the M class sample, and so on): $p_{\mathrm{X}} \approx 0.13\%$, $p_{\mathrm{M}} \approx 3.21\%$, $p_{\mathrm{C}} \approx 18.08\%$, $p_{\mathrm{NO1}} \approx 45.94\%$, $p_{\mathrm{NO2}} \approx 3.57\%$, $p_{\mathrm{NO3}} \approx 12.06\%$, $p_{\mathrm{NO4}} \approx 17.01\%$.
- Parsimony: We wanted each subset of samples to come from as few ARs as possible. In this way, we promoted training, validation, and test sets to be independent from each other, in the sense that samples belonging to the same AR must fall into the same data set.

2.3. Procedure for data sets generation

We set the size of the data set we wanted to create equal to n. This implied that we needed $n_{\mathrm{X}} = n \cdot p_{\mathrm{X}}$ samples labeled with X, $n_{\mathrm{M}} = n \cdot p_{\mathrm{M}}$ samples labeled with M, and so on. In terms of the procedure, first we randomly took an AR containing X flares and put all the samples from this AR into our data set. We continued to include new ARs with X flares until the number of samples labeled with X became n_{X}. If an AR contained more X flares than needed, we discarded those in excess. Next, we checked the amount of M flares in the data set under construction. If we had more than n_{M} samples with M flares, we randomly discarded the excess samples. If we had fewer than n_{M} samples with M flares, we randomly included ARs containing M flares (but not X flares) up to the correct rate. The process continued until we had the prescribed number of samples for each type.

In case more null class samples were needed, they could be added by randomly taking ARs that did not contain X and M flares, since the ones containing X and M flares were preserved for the construction of other independent and well-balanced data sets according to the parsimony and proportionality criterion, respectively.

We point out that the obtained algorithm is suboptimal since it may not find the smallest number of ARs needed, but it allowed a wide variety of independent data sets to be constructed as it operates randomly.

2.4. Validation of the forecasting algorithm

Each machine learning algorithm depends on certain parameters (e.g., weights utilized by each neuron of the neural network), which must be optimized on the basis of the historical data set. As such, in our experiments, the performance of the algorithm was evaluated on a validation set at the end of each epoch during the training phase. Then, the weight values corresponding to the best validation score were employed in the test phase.

2.5. Assessment of results

In order to compare the performances of different machine learning methods in flare forecasting, the following needed to be accounted for.

The classification results needed to be evaluated by considering appropriate skill scores defined on the so-called confusion matrix, which is characterized by four elements: true positives (TPs), that is to say, the number of samples labeled with YES and correctly predicted as positive; true negatives (TNs), meaning the number of samples labeled with NO and correctly predicted as negative; false positives (FPs), that is, the number of samples labeled with NO and incorrectly predicted as positive; and false negatives (FNs), the number of samples labeled with YES and incorrectly predicted as negative. In solar flare forecasting, the most meaningful skill scores are the ones specific for imbalanced data classification. Indeed, solar events are relatively seldom, as already pointed out in Sect. 2.1. Therefore, a chosen score needs to be able to represent the performance of the classifier concerning data sets with a small number of positive events. Among all possible skill scores, the true skill statistic (TSS; Hanssen & Kuipers 1965) is defined as

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

and its values have range in the interval $[-1, 1]$: when $\mathrm{TSS} = 1$, the performance is optimal, while $\mathrm{TSS} > 0$ means that the rates of positive and negative events are mixed up. The TSS is insensible to the class-imbalance ratio (Bloomfield et al. 2012), and therefore this is the skill score that we adopted in the present study.

Furthermore, the strategy outlined throughout Sects. 2.1–2.3 needed to be repeated several times in order to achieve some statistical significance. Therefore, many classification tests had to be carried out by generating different triples of training, validation, and test sets by randomly extracting AR magnetograms from the HMI archive.

Once results were obtained, some statistical indicators such as the mean value, the standard deviation value, the maximum value, and the minimum value needed to be reported. Obviously, the results achieved on the test set were not to be produced by applying any validation procedure directly to the test set.

3. Video data preparation

In general, the archive of the SDO/HMI mission includes 2D magnetograms of continuous intensity, of the full three-component magnetic field vector, and of the line-of-sight magnetic intensity. In the present study, we consider the Near Realtime Space Weather HMI Archive Patch (SHARP) data products associated with the line-of-sight components in the time range between 2012 September 14 and 2017 September 30. More specifically, our data products were 24-h-long videos made of 40 SHARP images of an AR, with 36 minutes cadence. This cadence is, in fact, a good trade-off to reduce the computational cost without loosing the meaningful information over time. Each image in these videos has been resized to 128 × 128 pixels, following similar procedures carried out by Huang et al. (2018) and Li et al. (2020), and based on bilinear interpolation. This size guarantees a good trade-off between computational efficiency and the performance reliability of the CNN (Bhattacharjee et al. 2020).

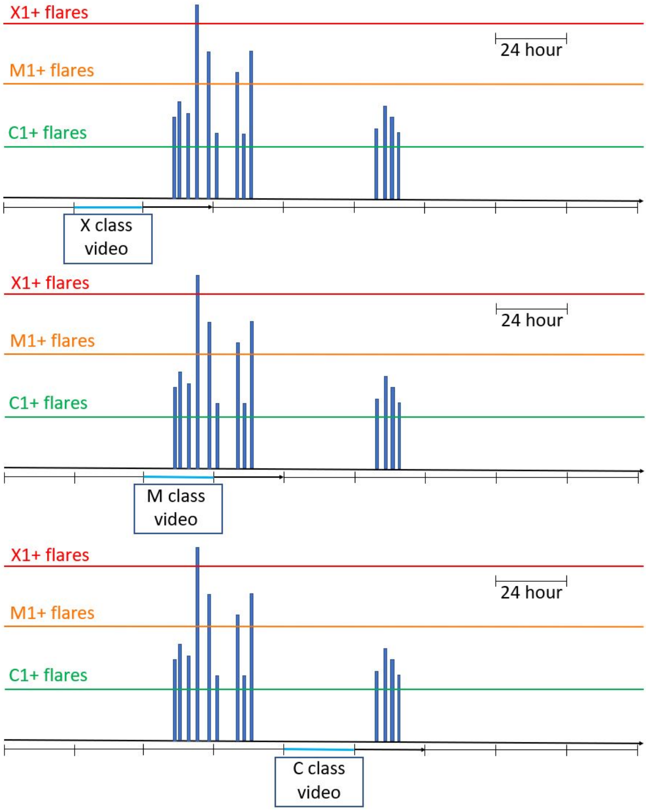

Fig. 1. *From top to bottom*: examples of X class, M class and C class videos.

Figures 1 and 2 are iconographical representations of how these videos are categorized with respect to the definitions given in Sect. 2. In particular, Fig. 1 illustrates a typical temporal history (from its birth to its death) of an AR that originates X1+, M1+, and C1+ flares (we point out that when the aim is to predict C1+ flares then the three kinds of videos are labeled with 1, whereas when the aim is to predict M1+ flares then just the first two kinds of videos are labeled with 1). Figure 2 provides schematic examples of NO2, NO3, and NO4 videos. NO1 data samples are not included in the figure, since they correspond to ARs that never originated a flaring event.

4. Deep learning method for HMI videos

The analysis of the video data samples was performed by means of an LRCN, which is a mixed deep learning model made of a CNN and an LSTM network. The first part of the LRCN network was made of the following sequence of layers (see Fig. 3, top panel): a 7×7 convolutional layer of 32 units; a 2×2 max-pooling layer; a 5×5 convolutional layer of 32 units; a 2×2 max-pooling layer; a 3×3 convolutional layer of 32 units; a 2×2 max-pooling layer; a 3×3 convolutional layer of 32 units; a 2×2 max-pooling layer; a dense layer of 64 units, where dropout was applied with a fraction of 0.1 input units dropped. Height and width strides were set to 2 for the convolutional layers and to 1 for the max-pooling. Each convolutional layer was L_2-regularized and the corresponding output was standardized. Before applying the dense layer, the last pooling layer was flattened. The Rectified Linear Unit (ReLU) was used as an activation function in all

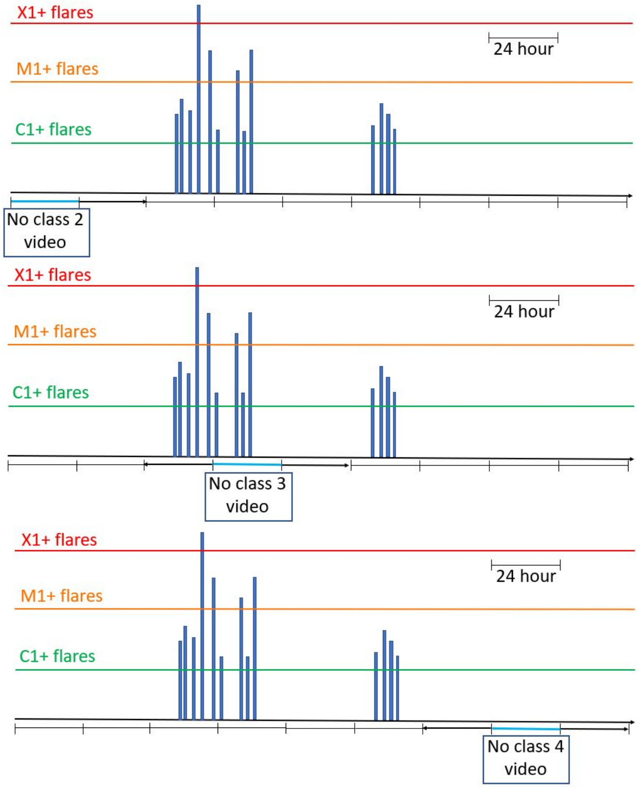

Fig. 2. *From top to bottom*: examples of NO2, NO3 and NO4 class videos.

layers. We also point out that the input videos, which consisted of 40 frames of 128 × 128 magnetograms each, were treated as time series, so that the CNN architecture described above was applied to each video frame in parallel.

In the second part of the LRCN, the outputs of the CNNs (40 vectors, each one composed by 64 features) were sequentially considered in time and then passed to the LSTM (see Fig. 3, bottom panel), which consisted of 50 units. Similarly to the dense layer, here dropout was applied with a fraction of 0.5 active units. Finally, a dense sigmoid unit drove the output of the LSTM to be in the interval [0, 1], in order to perform binary classification.

The CNN-LSTM network was trained for 100 epochs by taking batches of 128 samples. Moreover, the *Adam* scheme (Kingma & Ba 2015) was adopted for the optimization of the weights.

As far as the loss function utilized in the training phase is concerned, in the present study we propose using a score-oriented loss (SOL) function (Marchetti et al. 2021), which allows an automated optimization of a given skill score without the need of an a posteriori choice of the optimal threshold that converts the probabilistic outcomes into binary classification. Specifically, the SOL function applied in this paper is based on the optimization of the TSS, which is highly insensitive to the class imbalance ratio in the training set. The realization of this score-driven strategy was performed as follows.

The classical confusion matrix depends on a fixed threshold parameter $\tau \in (0, 1)$, meaning that,

$$\mathrm{CM}(\tau) = \begin{pmatrix} \mathrm{TN}(\tau) & \mathrm{FP}(\tau) \\ \mathrm{FN}(\tau) & \mathrm{TP}(\tau) \end{pmatrix}. \qquad (2)$$

For the construction of SOL functions, the threshold parameter τ is dealt with as a random variable associated with a specific probability density function. Letting $\mathbb{E}_\tau[\cdot]$ be the expected value with respect to τ, we took an expected confusion matrix

$$\mathbb{E}_\tau[\mathrm{CM}(\tau)] = \begin{pmatrix} \mathbb{E}_\tau[\mathrm{TN}(\tau)] & \mathbb{E}_\tau[\mathrm{FP}(\tau)] \\ \mathbb{E}_\tau[\mathrm{FN}(\tau)] & \mathbb{E}_\tau[\mathrm{TP}(\tau)] \end{pmatrix}. \qquad (3)$$

From this matrix it was possible to construct the expected TSS

$$\mathbb{E}_\tau[\mathrm{TSS}(\tau)] = \frac{\mathbb{E}_\tau[\mathrm{TP}(\tau)]}{\mathbb{E}_\tau[\mathrm{TP}(\tau) + \mathrm{FN}(\tau)]} - \frac{\mathbb{E}_\tau[\mathrm{FP}(\tau)]}{\mathbb{E}_\tau[\mathrm{FP}(\tau) + \mathrm{TN}(\tau)]} - 1, \qquad (4)$$

and from this the TSS-driven loss function

$$\ell_{\mathrm{TSS}} := -\mathbb{E}_\tau[\mathrm{TSS}(\tau)]. \qquad (5)$$

This function is differentiable, and therefore can be easily minimized in the training phase. It has the crucial advantage that the corresponding skill score is automatically optimized, without the need of any a posteriori tuning of the thresholding parameter τ, which is set to the default value 0.5.

5. Results

The LRCN described in the previous section was applied to video data generated as described in Sect. 3, using the validation strategy illustrated in Sect. 2. We first filled up a training set, a validation set, and a test set made of 3000, 750, and 750 data samples, respectively, and we used data augmentation to increase the cardinality of these sets up to 15 000, 3750, and 3750, respectively. The data augmentation process here relied on rotating magnetograms by 90 and 180 degrees, and flipping magnetograms horizontally and vertically as done by Li et al. (2020). We repeated this set generation process ten times in order to create ten random realizations of these three sets. Table 2 shows the prediction results obtained by the LRCN in the case of the realization of the validation and test sets. Specifically, the table focuses on the TSS and provides its mean and standard deviation values, the minimum value, the 25th percentile value, the median value, the 75th percentile value and the maximum value for the prediction of both C1+ and M1+ flares. The numbers in this table show that image-based deep learning is more effective at predicting M1+ flares than C1+ flares, in line with most results in the scientific literature. Indeed, from a heuristic viewpoint, the reason for this could be the fact that M1+ flares better distance themselves from the null event cases. Furthermore, both the mean and the median values for the test sets are close to the ones provided by the network when applied to the validation sets, and in all cases the standard deviations are nicely small.

Figures 4 and 5 aim at providing a quantitative confirmation of the implementation paradigm introduced in Sect. 2. The boxplots in Fig. 4 represent the rates of the TPs and TNs computed on the ten random realizations of the test set. The rates of the TPs and TNs correspond to the standard true positive rate (also known as sensitivity) and the true negative rate (also known as specificity), respectively. Using rates instead of TP and TN units allows us to evaluate the performance of the method over ten different datasets. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median and a star at the mean, while the whiskers extend to the range of the data, but no more than 1.5(Q3–Q1) from the edges of the box. Outliers are plotted as separate dots. The boxplots show that: when the aim is to predict C1+ flares, X class and M class flares are predicted with a higher success rate with respect to C flares, coherent with the fact that ARs labeled as either M or X are more distinguishable from ARs associated with NO-class flares (this is particularly true for all cases where flares are close to strong B events).

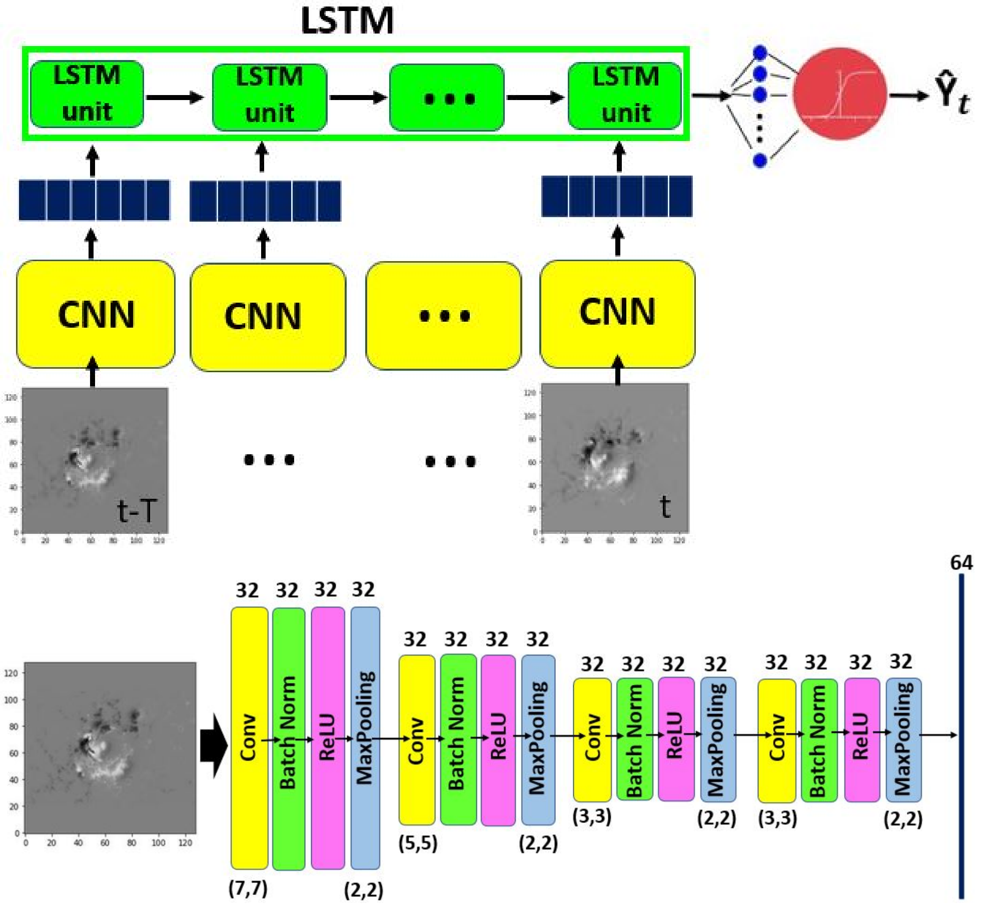

Fig. 3. *Top panel*: overall LRCN design. *Bottom panel*: CNN architecture.

Table 2. Mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values of the TSS distribution computed on 10 validation sets and 10 test sets, separately.

	TSS (C1+ flares)						
	Mean	Std	Min	25th perc	Median	75th perc	Max
Validation	0.57	0.02	0.55	0.56	0.57	0.59	0.61
test	0.55	0.05	0.46	0.52	0.54	0.60	0.61
	TSS (M1+ flares)						
	Mean	Std	Min	25th perc	Median	75th perc	Max
Validation	0.76	0.07	0.65	0.67	0.77	0.82	0.85
test	0.68	0.09	0.55	0.61	0.69	0.72	0.82

For X flare predictions, the error bar is zero as the network correctly hits the X flare prediction all ten times. The easiest null events to predict are those where no activity is reported in the entire AR history (NO1). Then, NO2 and NO4 cases are similar and are associated with situations distinctly separated from C, M, and X samples. NO3 samples involve time windows during which the emission is occurring, and so they can be more easily confused with C, M, and X samples. When the aim is to predict M1+ flares, the smallest TN rate refers to C class data samples, consistent with the fact that such videos may be associated with events that release an amount of energy close to that of weak M flares.

In Fig. 5, we computed the TSS values while successively excluding data samples of different classes from the test sets. We found that, when predicting C1+ events, the TSS values are smallest when all classes are represented in the test sets, and nicely increase while successively and cumulatively excluding NO2, NO3, and NO4 events. On the other hand, when predicting M1+ flares, the TSS values significantly increase when data samples belonging to the C class are excluded from the test sets.

6. Comments and conclusions

The scientific rationale of the present study was twofold. First, we aimed to verify the feasibility of a fully automated flare forecasting procedure that takes videos of line-of-sight magnetograms as input and provides binary predictions as output. Second, we also aimed to study the impact of the data set preparation on deep learning performances. The first conclusion we can draw from this analysis is that deep learning is able to realize flare video classification with prediction performances that are in line with the ones obtained by machine learning approaches that require an a priori extraction of features from the

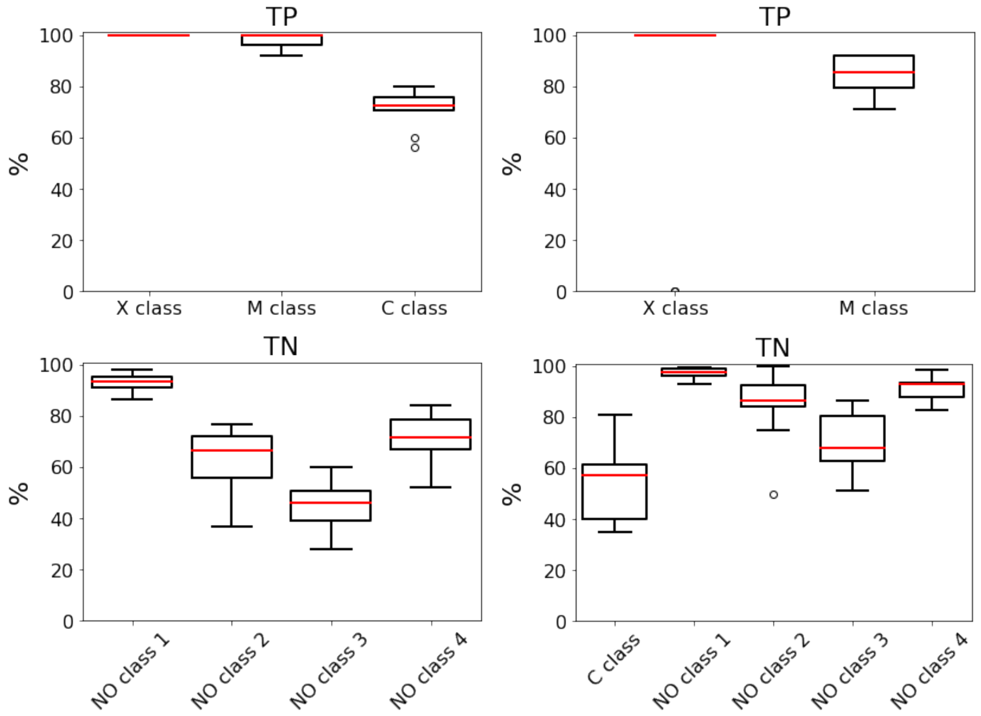

Fig. 4. Boxplots of the percentages of correctly predicted samples in the10 test sets for the C1+ flares prediction (*first column*) and for the M1+ flares prediction (*second column*). The rates of the TPs and the rates of the TNs are shown in the *top and bottom panels*, respectively.

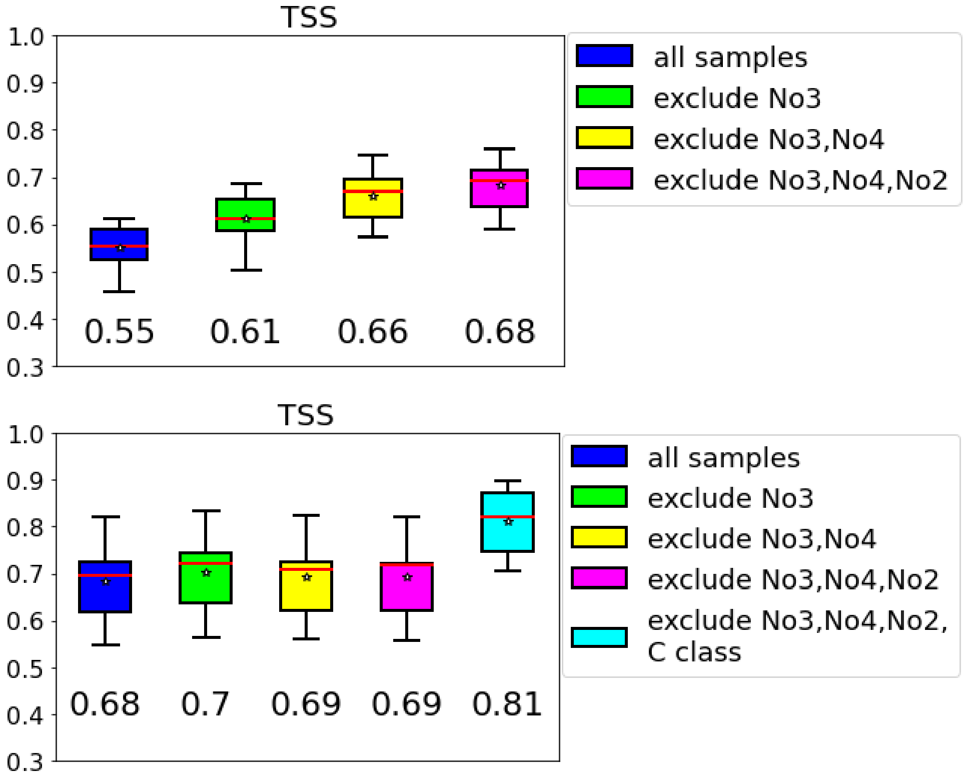

Fig. 5. Boxplots of TSS values computed on 10 test sets for the prediction of C1+ class flares (*top panel*) and M1+ class flares (*bottom panel*) by successively excluding data samples of different classes from the test sets. *Top panel*. From left to right: (1) all classes are present in the test sets; (2) NO3 samples are excluded from the test sets; (3) NO3 and NO4 samples are excluded from the test sets; (4) NO3, NO4, and NO2 samples are excluded from the test sets. *Bottom panel*. From left to right: (1) all classes are present in the test sets; (2) NO3 samples are excluded from the test sets; (3) NO3 and NO4 samples are excluded from the test sets; (4) NO3, NO4, and NO2 samples are excluded from the test sets; (5) NO3, NO4, NO2, and C class samples are excluded from the test sets.

HMI magnetograms, by means of pattern recognition and feature computation algorithms. Furthermore, in our implementation of the convolutional network, the use of a SOL function allows an a priori optimization of the TSS, which implies avoiding the application of (often critical) a posteriori thresholding on the score value. From a computational viewpoint, the most demanding part of this video-based procedure is the training phase. The training process on one data set takes about 2 h on a machine NVIDIA DGX deep learning workstation with 4 GPUs Tesla V100-32GB, which makes the procedure operationally feasible, considering the fact that the training and validation phases can be precalculated.

Second, the results in Fig. 4 and in Fig. 5 clearly show that the way the training and validation sets are prepared for the network optimization has a really significant impact on the prediction performances. In particular, these figures prove that an appropriate balancing of these sets should account not only for the presence of ARs generating flares, but also for the presence of ARs associated with null events of different kinds. In fact, the use of databases with the same occurrence rate is the theoretical starting point in machine learning foundations, when one wants to assess the prediction effectiveness of different algorithms. Therefore, in this paper we have averaged the impact of the temporal aspect, comparing algorithms in a setting where the distortions due to solar cycle are negligible. Furthermore, the nice aspect of our approach is that it can be readily specialized to the operational setting by modifying the database generation procedure, in order to account for the different flare occurrence rates characteristic of a specific temporal range. Also, this implementation paradigm can be used even when the input data sets are made of images, extracted features, and time series of an extracted feature.

We further point out that the TSS values obtained in this analysis are distinctively different from 1, as typically occurs in most flare forecasting studies based on machine learning. This video-based study fully exploits the dynamical information contained in HMI magnetograms. Yet, it provides TSS values comparable with those obtained with point-in-time, feature-based approaches. From an astrophysical viewpoint, this finding can be interpreted by referring to the fact that HMI data do not contain information on the low solar corona, where flares are generated, but just on the photospheric layer. A further interpretation might rely on the fact that flares can be modeled within stochastic frameworks, which would hamper the possibility of binary predictions, in favor of probabilistic indications of flare occurrence (Rosner & Vaiana 1978; Wheatland & Litvinenko 2002; Aschwanden et al. 2016; Campi et al. 2019). In this perspective, two possible further lines of research are: the combination of image-based features with features that rely on noniconographic information within fully data-driven models; and the exploitation of physical information in the design and optimization of the networks.

As a final remark, we can probably state that a robust comparison of skill scores obtained in recent supervised flare forecasting studies may be significantly compromised by the use of training, validation, and test sets that are not generated according to a shared process. In particular, in most papers there has been a trend to chase high skill scores without an adequate focus on data preparation. This means that issues such as those related to concept drift (Žliobaitė et al. 2016), that is to say, the fact that data evolve over time, have seldom been acknowledged. On the other hand, a common trend for most supervised methods is that they obtain higher skill scores while predicting M flares (this is true also for our deep neural network). The reason for this behavior is probably related to the fact that M flares stand out from null events more than C flares.

Acknowledgements. The authors kindly acknowledge the financial contribution from the AI-FLARES ASI-INAF n.2018-16-HH.0 agreement.

References

Ahmadzadeh, A., Aydin, B., Georgoulis, M. K., et al. 2021, ApJS, 254, 23
Aschwanden, M. J. 2008, J. Astrophys. Astron., 29, 3
Aschwanden, M. J., Crosby, N. B., Dimitropoulou, M., et al. 2016, Space Sci. Rev., 198, 47
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016a, ApJ, 829, 89
Barnes, G., Leka, K., Schrijver, C., et al. 2016b, ApJ, 829, 89
Bhattacharjee, S., Alshehhi, R., Dhuri, D. B., & Hanasoge, S. M. 2020, ApJ, 898, 98
Bloomfield, D. S., Higgins, P. A., McAteer, R. J., & Gallagher, P. T. 2012, ApJ, 747, L41
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Campi, C., Benvenuto, F., Massone, A. M., et al. 2019, ApJ, 883, 150
Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, Space Weather, 17, 1404
Crown, M. D. 2012, Space Weather, 10, 6
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, Sol. Phys., 293, 1
Georgoulis, M. K., Bloomfield, D. S., Piana, M., et al. 2021a, J. Space Weathe Space Climate, 11, 39
Georgoulis, M. K., Martens, P., Aydin, B., et al. 2021b, in 43rd COSPAR Scientific Assembly. Held 28 January - 4 February, 43, 2357
Hanssen, A., & Kuipers, W. 1965, On the Relationship Between the Frequency of Rain and Various Meteorological Parameters: (with Reference to the Problem Ob Objective Forecasting), Koninkl. Nederlands Meterologisch Institut. Mededelingen en Verhandelingen (Staatsdrukkerij- en Uitgeverijbedrijf)
Hochreiter, S., & Schmidhuber, J. 1997, Neural Comput., 9, 1735
Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
Hudson, H. S. 2011, Space Sci. Rev., 158, 5
Jonas, E., Bobra, M., Shankar, V., Hoeksema, J. T., & Recht, B. 2018, Sol. Phys., 293, 1
Kingma, D. P., & Ba, J. 2015, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, eds. Y. Bengio, & Y. LeCun, Conf. Track Proc.
Leka, K. D., Park, S.-H., Kusano, K., et al. 2019a, ApJS, 243, 36
Leka, K. D., Park, S.-H., Kusano, K., et al. 2019b, ApJ, 881, 101
Li, X., Zheng, Y., Wang, X., & Wang, L. 2020, ApJ, 891, 10
Liu, C., Deng, N., Wang, J. T., & Wang, H. 2017, ApJ, 843, 104
Liu, H., Liu, C., Wang, J. T., & Wang, H. 2019, ApJ, 877, 121
Marchetti, F., Guastavino, S., Piana, M., & Campi, C. 2021, ArXiv preprint [arXiv:2103.15522]
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
McAteer, R. T. J., Gallagher, P. T., & Conlon, P. A. 2010, Adv. Space Res., 45, 1067
Murray, S. A., Bingham, S., Sharpe, M., & Jackson, D. R. 2017, Space Weather, 15, 577
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. 2018, ApJ, 858, 113
Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., & Ishii, M. 2020, ApJ, 899, 150
Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., & Ishii, M. 2021, Earth Planets Space, 73, 64
Park, E., Moon, Y.-J., Shin, S., et al. 2018, ApJ, 869, 91
Park, S.-H., Leka, K., Kusano, K., et al. 2020, ApJ, 890, 124
Petrakou, E. 2018, J. Atm. Sol. Terr. Phys., 175, 18
Rosner, R., & Vaiana, G. S. 1978, ApJ, 222, 1104
Schwenn, R. 2006, Liv. Rev. Sol. Phys., 3, 1
Shibata, K. 1996, Adv. Space Res., 17, 9
Shibata, K., & Magara, T. 2011, Liv. Rev. Sol. Phys., 8, 6
Song, H., Tan, C., Jing, J., et al. 2009, Sol. Phys., 254, 101
Strugarek, A., & Charbonneau, P. 2014, Sol. Phys., 289, 4137
Su, Y., Veronig, A. M., Holman, G. D., et al. 2013, Nat. Phys., 9, 489
Sui, L., Holman, G. D., & Dennis, B. R. 2004, ApJ, 612, 546
Tandberg-Hanssen, E., & Emslie, A. G. 1988, The Physics of Solar Flares
Vapnik, V. 1998, Statistical Learning Theory (Wiley)
Wang, X., Chen, Y., Toth, G., et al. 2020, ApJ, 895, 3
Wheatland, M. S., & Litvinenko, Y. E. 2002, Sol. Phys., 211, 255
Yi, K., Moon, Y.-J., Lim, D., Park, E., & Lee, H. 2021, ApJ, 910, 8
Yu, Y., Si, X., Hu, C., & Zhang, J. 2019, Neural Comput., 31, 1235
Žliobaitė, I., Pechenizkiy, M., & Gama, J. 2016, Big data analysis: New Algorithms for a New Society, 91