The Astrophysical Journal, 883:150 (12pp), 2019 October 1 https://doi.org/10.3847/1538-4357/ab3c26

Feature Ranking of Active Region Source Properties in Solar Flare Forecasting and the Uncompromised Stochasticity of Flare Occurrence

Cristina Campi[1], Federico Benvenuto[2], Anna Maria Massone[2,3], D. Shaun Bloomfield[4], Manolis K. Georgoulis[5,6], and Michele Piana[2,3]

[1] Dipartimento di Matematica "Tullio Levi Civita," Università di Padova, via Trieste 63 I-35121 Padova, Italy; cristina.campi@unipd.it

[2] Dipartimento di Matematica, Università di Genova, via Dodecaneso 35 I-16146 Genova, Italy; benvenuto@dima.unige.it, massone@dima.unige.it, piana@dima.unige.it

[3] CNR - SPIN Genova, via Dodecaneso 33 I-16146 Genova, Italy

[4] Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle upon Tyne, NE1 8ST, UK; shaun.bloomfield@northumbria.ac.uk

[5] Department of Physics and Astronomy, Georgia State University, Atlanta (GA), USA; manolis.georgoulis@academyofathens.gr

[6] RCAAM of the Academy of Athens, Athens, Greece

Received 2019 June 27; revised 2019 August 12; accepted 2019 August 14; published 2019 September 30

Abstract

Solar flares originate from magnetically active regions (ARs) but not all solar ARs give rise to a flare. Therefore, the challenge of solar flare prediction benefits from an intelligent computational analysis of physics-based properties extracted from AR observables, most commonly line-of-sight or vector magnetograms of the active region photosphere. For the purpose of flare forecasting, this study utilizes an unprecedented 171 flare-predictive AR properties, mainly inferred by the Helioseismic and Magnetic Imager on board the *Solar Dynamics Observatory* (*SDO*/HMI) in the course of the European Union Horizon 2020 FLARECAST project. Using two different supervised machine-learning methods that allow feature ranking as a function of predictive capability, we show that (i) an objective training and testing process is paramount for the performance of every supervised machine-learning method; (ii) most properties include overlapping information and are therefore highly redundant for flare prediction; (iii) solar flare prediction is still—and will likely remain—a predominantly probabilistic challenge.

Unified Astronomy Thesaurus concepts: Astronomy data analysis (1858); Neural networks (1933); Solar flares (1496); Solar activity (1475); Solar active region magnetic fields (1975)

1. Introduction

Solar flares are the most explosive events in the heliosphere, releasing in an abrupt way up to 10^{33} erg of energy in a time interval typically ranging between 10 and 1000 s (Benz 2017). This energy is previously stored in specific magnetic configurations and, when magnetic reconnection occurs (Priest & Forbes 2002), it is transformed into mass acceleration, heating, and electromagnetic radiation at all wavelengths. It is also established that flares are a major space weather agent in the heliosphere (Schwenn 2006), while, as secondary effects through their correlation with coronal mass ejections, they induce geospace and ionospheric disturbances, malfunctions, and impairments on technologies in the geosphere, such as flight navigation, satellite communication, and power grid distribution.

Solar flare forecasting is a prominent discipline (Gallagher et al. 2002; Georgoulis & Rust 2007; Li et al. 2007, 2008; Schrijver 2007; Barnes & Leka 2008; Wang et al. 2008; Colak & Qahwaji 2009; Yu et al. 2009; Ahmed et al. 2013; Bobra & Couvidat 2015; Barnes et al. 2016; McCloskey et al. 2017; Murray et al. 2017; Sadykov & Kosovichev 2017; Benvenuto et al. 2018; Huang et al. 2018; Massone et al. 2018; Nishizuka et al. 2018; Park et al. 2018) within the recent field of space weather forecasting that relies on the availability of two ingredients; one observational and one computational. First, it is well-established that solar active regions (ARs) exclusively host major flares and therefore flare prediction needs experimental data on AR properties, associated to the photospheric and coronal magnetic field; however, coronal information has only recently started being used in the form of EUV images given as input to a deep learning network by Nishizuka et al. (2018). Second, this information on AR magnetic properties can be processed for prediction purposes by means of a computational method for data analysis; machine learning has recently offered strong candidates for such methods.

Since February 2010, the Helioseismic and Magnetic Imager on board the *Solar Dynamics Observatory* (*SDO*/HMI; Scherrer et al. 2012) is providing both line-of-sight and vector magnetograms of the full solar disk at a (vector magnetogram) cadence of 12 minutes. *SDO*/HMI magnetograms can be used for solar flare prediction according to two different approaches. First, HMI magnetograms are utilized to calculate a variety of properties either from the line-of-sight component only, from the radial component only, or from all three vector components. Various single-valued quantities, hereafter referred to as features, can be calculated from these property images through a variety of techniques (e.g., thresholding, feature recognition, etc.), such that calculation of one physical property may provide multiple features as inputs to machine learning (i.e., image maximum, total, and moments). Of course, additional features that are not derived from property images may also contribute to the input data set. Second, from a deep learning perspective, HMI images can be given as input to Convolutional Neural Networks that automatically perform a probabilistic forecasting. This present paper follows the first approach, and this is for several reasons. First, we had at our disposal the property extraction power provided by the algorithms developed within the Horizon 2020 FLARECAST project (http://flarecast.eu), which generated data sets of almost 200 features determined from properties extracted from photospheric *SDO*/

HMI vector magnetograms. This database of features probably represents the highest data dimensionality currently available for flare forecasting purposes. Second, one of the objectives of our research was to determine to what extent AR properties are redundant when forecasting flares, and a straightforward way to do this is by ranking the extracted features according to their predictive capability. Finally, so far most publications in flare prediction utilize feature-based machine-learning methods, so another objective of this paper is to investigate how data preparation (and, specifically, the preparation of the training set in the case of supervised algorithms) impacts the prediction scores. Specifically, the analysis performed in this paper relies on two supervised machine-learning algorithms that combine prediction with feature ranking, namely hybrid LASSO (Benvenuto et al. 2018) and Random Forest (RF; Breiman 2001). However, two other methods of this kind, namely logit (Wu et al. 2009) and a support vector machine for classification (Cortes & Vapnik 1995), have been also applied to the same data sets for verification purposes, with coherent results.

The content of the paper is as follows. Section 2 overviews the data analysis procedure, describing in detail the features used for prediction, the data preparation process, and key aspects of the machine-learning methods adopted. Section 3 contains the results of the analysis, while Section 4 discusses these results. Our conclusions are offered in Section 5.

2. Methods

2.1. Data and Features

Our analysis relies on the Near-Realtime Space Weather HMI Archive Patch (SHARP) data product of the HMI database (Bobra et al. 2014). These data comprise 2D images of continuum intensity, the full three-component magnetic field vector, and the line-of-sight component of each photospheric HARP. We then made use of property extraction algorithms developed by the FLARECAST Consortium in order to construct a property database made of property vectors comprising up to 171 components. FLARECAST algorithms first extracted the following 167 features, often duplicating the property calculation step on B_{los} and B_{radial} input data, as per the findings of Guerra et al. (2018):

1. Schrijver's R-value (Schrijver 2007): one property yielding a total of two features.
2. Multifractal structure and function spectrum on a 2D image: two properties yielding a total of four features.
3. Falconer's total free magnetic energy proxy WL_{SG} (Falconer et al. 2008): one property yielding a total of two features.
4. Sum of the horizontal magnetic gradient, G_S, and the separation of opposite-polarity sunspot subgroups, S_{l-f} (Korsos & Erdelyi 2016): two properties yielding a total of four features.
5. Spectral power indices extracted by means of the Fourier transform and of a continuous wavelet transform: one property yielding a total of four features.
6. Magnetic polarity inversion line (MPIL) characteristics: three properties yielding a total of six features.
7. Effective connected magnetic field strength (B_{eff}): one property yielding a total of two features.
8. Vertical decay index of potential field: four properties yielding a total of eight features.
9. Nonneutralized electric currents: one property yielding one feature.
10. Ising energy (E): one property yielding a total of four features.
11. Fractal dimension (D): one property yielding a total of two features.
12. Flow field characteristics: six properties yielding a total of 16 features.
13. Magnetic helicity and energy injection rate: 14 properties yielding a total of 14 features.
14. SHARP keywords calculated from their corresponding vector and line-of-sight magnetograms: 16 properties yielding a total of 100 features (including the maximum, total, median, mean, standard deviation, skewness and kurtosis over the SHARP field-of-view).

SHARP ARs are associated to solar flares of *GOES* class C1 and above (C1+) and solar flares of *GOES* class M1 and above (M1+) by means of a standard procedure. It is first verified whether the SHARP data contain NOAA-numbered regions (i.e., sunspot groups) by comparison with NOAA's daily Solar Region Summary (SRS) file immediately before the SHARP observations. Then, if any NOAA number is assigned to the SHARP data, the process searches the NOAA/SWPC daily events lists for *GOES* flares occurring in the same source region during the entire disk passage. Once the flare association is realized, the following four details become available for all flares and these are used in assigning flare outcome labels:

1. *GOES* peak magnitude (F_M).
2. Time difference (in seconds) between the SHARP observation time and the flare start time (τ_s).
3. Time difference (in seconds) between the SHARP observation time and the flare peak time (τ_p).
4. Time difference (in seconds) between the SHARP observation time and the flare end time (τ_e).

Eventually, this analysis provides 167 features extracted from the HMI images. Four further features come from the NOAA/SRS database: the mean heliographic longitude and latitude of each AR, a binary label encoding the presence of a flare in the past 24 hr and the flare index of events occurring within the past 24 hr. A summary of all resulting features considered in the analysis can be found at both the url https://api.flarecast.eu/property/ui/ and in Tables 3–5 in the Appendix.

2.2. Data Preparation

The experiment designed in this paper relies on supervised machine learning, which requires appropriate historical sets to train the prediction networks. To enforce consistency in time and robustness of our tests, we constructed four training sets, each one corresponding to a specific forecast issuing time expressed as universal time [UT], namely 00:00, 06:00, 12:00, and 18:00. For each issuing time we considered the set of *SDO*/HMI images recorded at that time in the range of days between 2012 September 14 and 2016 April 30, with 24 hr sampling. While filling up the training set we took care to focus on ARs rather than on feature vectors. In fact, around two-thirds of ARs were randomly extracted from the set of all ARs belonging to a specific issuing time and the 171-dimensional feature vectors associated to each AR were labeled by annotating whether a *GOES* C1+ flare occurred in the next 24 hr. The set of feature vectors associated to the remaining one-third of ARs was not labeled and was provided as a test set for experiments on supervised learning algorithms trained on the training set. In this manner, training and testing do not overlap in any way, either in time or in terms of ARs examined. We finally point out that, for each issuing time, the random, complete separation of ARs into training and test sets was replicated 100 times to enable statistical robustness of the results. A similar procedure was implemented to generate training sets to use for the prediction of *GOES* M1+ flares. The reason why we did not consider the prediction of flares with class above X is because they are extremely seldom in the database (less than 0.2% of the overall point-in-time events in the original training set). C1+ and M1+ flares are around 26% and 4% of the set content, respectively.

2.3. Prediction and Feature Ranking Methods

Two different machine-learning methods, namely hybrid LASSO (HLA) and RF are utilized in this paper, for both performing the binary prediction of the flare occurrence and for additionally identifying the effectiveness with which the different features contribute to the prediction.

LASSO methods (Tibshirani 1996) are intrinsically regression methods and therefore they are not originally conceived for applications that require a binary YES/NO output. However, in Benvenuto et al. (2018) a threshold optimization is introduced to the LASSO outcome in order to realize classification by means of fuzzy clustering (Bezdek et al. 1984). The idea of HLA is therefore to use LASSO in the first step in order to promote sparsity and to realize feature selection; this step provides an optimal estimate of the model parameters and corresponding predicted output. In the second step, Fuzzy C-Means is applied for clustering the predicted output in two classes. The main advantage of this approach is in the use of fuzzy clustering to automatically classify the regression output in two classes. Indeed, fuzzy clustering identifies flaring/nonflaring events with a thresholding procedure that is data-adaptive and completely operator-independent. Details about HLA as implemented in the present paper can be found in Benvenuto et al. (2018).

RF belongs to the family of the ensemble methods, i.e., methods that make use of a combination of different learning models to increase the classification accuracy. In particular, RF works as a large collection of decorrelated decision trees. In fact, here the training set is randomly divided into 10 subsets and for each subset a separate decision tree is built. Each decision tree is then used to classify an incoming unlabeled sample. If correctly implemented, RF can be used as feature rankers. In fact, the relative depth of a feature used as a decision node in a tree can be identified as the relative importance of that feature with respect to the predictability of the target variable. Features used at the top of the tree contribute to the final prediction decision of a larger fraction of the input samples. The expected fraction of the samples they contribute to can thus be used as an estimate of the relative importance of the features. Details about RF as implemented in the present paper can be found in Breiman (2001).

Once the two machine-learning methods have been applied to the input data, predictors are ranked by using recursive feature elimination (RFE). This iterative procedure can be summarized as follows:

1. Train the classifier.
2. Compute the ranking for all features.
3. Remove the feature with the smallest ranking.

Details about RFE as implemented in the present paper can be found in Guyon et al. (2002).

3. Results

The effectiveness of the prediction was assessed by skill scores computed on the previously unseen test sets. Following suggestions in Bloomfield et al. (2012), we chose to use the true skill statistic (TSS) and the Heidke skill score (HSS), assuming them as representative among skill scores existing in the literature. This said, although not shown here, we performed the analysis using the false alarm ratio, probability of detection, and accuracy metrics, obtaining similar results in terms of the relative forecasting effectiveness of the two machine-learning methods. We point out that all these scores are computed by means of binary predictions applied to the test set. However, as noted in Section 2.3, LASSO and RF are regression methods providing as outcome real positive numbers that can be interpreted in a probabilistic sense and, in our approach, the transformation of these variables into dichotomous yes/no responses is accomplished by applying a fuzzy clustering technique against the regression outcomes. Other works typically apply an arbitrary probability threshold, P_{th}, of 0.5 to create dichotomous forecasts (Leka et al. 2019a, 2019b), although it should be noted that discriminating thresholds optimized on TSS should find P_{th} values close to the climatology (i.e., the average flare-day rate; Bloomfield et al. 2012; Barnes et al. 2016). Figure 1 shows a comparison between the thresholding performances of the clustering technique and the ones provided by the optimization of TSS and HSS and by the use of an ROC curve. Interestingly, the two hybrid regression/clustering approaches provide similar results, which are rather conservative and rather close to the ones achieved by optimizing the HSS, especially in the case of the prediction of C1+ flares. Furthermore, the ROC curve method relies on cutoff values computed by means of the Youden index (Youden 1950), which formally leads to the maximization of the TSS; in fact the figure shows that the two values are always very close and the small differences are just related to the different numerical way the thresholding-search schemes were implemented. For C1+ flares, our hybrid approach results in $P_{\text{th}} \approx 0.4$ for both HLA and RF, meaning that our C1+ TSS values are (probably) more comparable to those whose probabilities are converted to dichotomous forecasts using $P_{\text{th}} = 0.5$ (as our P_{th} lies closer to 0.5 than the average C1+ flare-day rate of ~0.26). The situation is more complex for M1+ flares, however, as the average P_{th} found by the fuzzy-clustered HLA method is almost equivalent to that optimized on TSS (i.e., approaching the average flare-day rate of 0.04) while for the fuzzy-clustered RF method it instead occupies greater values that lie between the TSS and HSS optimized cases.

The averages and standard deviations of the TSS and HSS values over 100 random realizations of the training/test sets for both prediction methods are shown in Table 1. In the case of HSS, the reliability of average values may be challenged by inconsistent flare/no-flare imbalance ratio across the 100 realizations. However, we have a posteriori checked the sample statistics of the 100 random realizations: average flare/no-flare imbalance ratios across the 100 test sets are ~0.34 for C1+ events and ~0.04 for M1+ events, with relative standard deviations that are <16% and <27% of these values, respectively (reflecting the

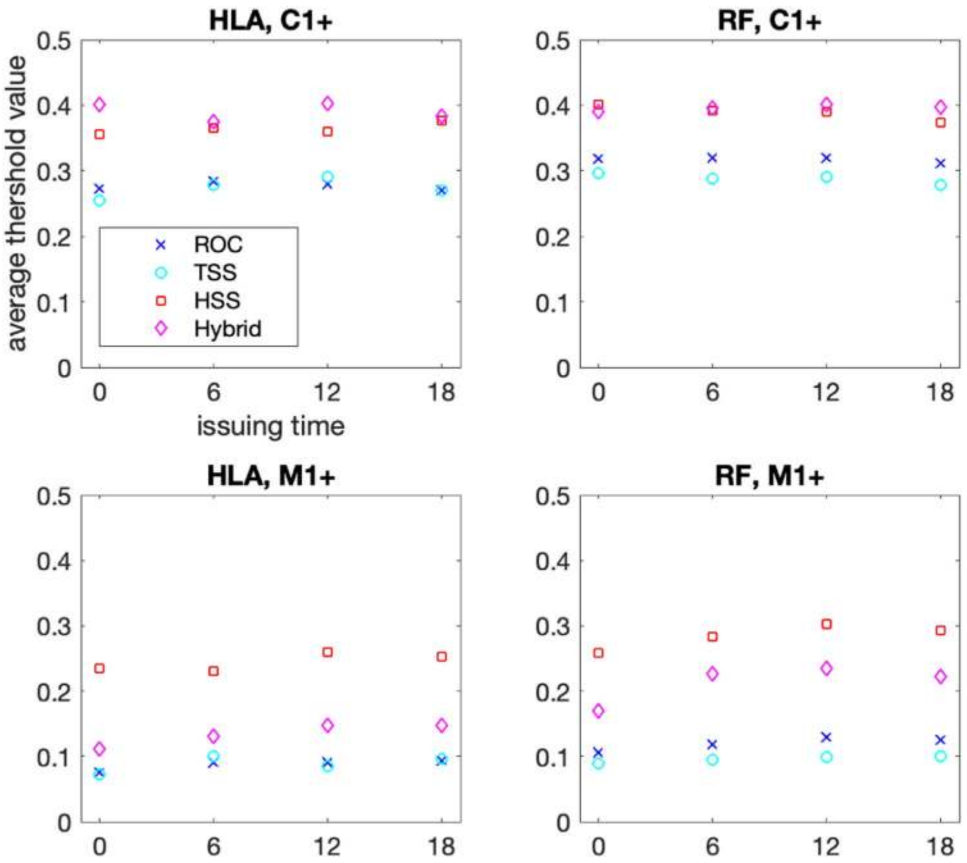

Figure 1. Probability threshold values, P_{th}, averaged over the 100 realizations of the training set. In each panel, symbols indicate the approach applied: hybrid fuzzy clustering (diamonds); HSS optimization (squares); TSS optimization (circles); ROC curve YoudenÕs index optimization (crosses). Top row: prediction of C1+ flares with LASSO and RF (left and right panels, respectively). Bottom row: prediction of M1+flares with LASSO and RF (left and right panels, respectively).

Table 1
Average TSS- and HSS-values, Along with Applicable Standard Deviations, Over the Outcomes of HLA and RF as Applied Against 100 Random Realizations of the Training/test Sets

	Test Set-C1+	Test Set-C1+	Test Set-M1+	Test Set-M1+
00:00:00UT	TSS	HSS	TSS	HSS
HLA	0.48 ± 0.06	0.51 ± 0.05	0.56 ± 0.14	0.27 ± 0.06
RF	0.53 ± 0.05	0.52 ± 0.04	0.48 ± 0.14	0.33 ± 0.09
06:00:00UT	TSS	HSS	TSS	HSS
HLA	0.53 ± 0.03	0.54 ± 0.03	0.67 ± 0.05	0.35 ± 0.04
RF	0.54 ± 0.03	0.54 ± 0.03	0.49 ± 0.08	0.42 ± 0.06
12:00:00UT	TSS	HSS	TSS	HSS
HLA	0.51 ± 0.04	0.54 ± 0.03	0.66 ± 0.06	0.38 ± 0.04
RF	0.53 ± 0.03	0.53 ± 0.03	0.51 ± 0.09	0.43 ± 0.06
18:00:00UT	TSS	HSS	TSS	HSS
HLA	0.54 ± 0.04	0.55 ± 0.03	0.64 ± 0.07	0.39 ± 0.04
RF	0.55 ± 0.03	0.55 ± 0.03	0.53 ± 0.09	0.43 ± 0.06

largest relative standard deviations for the four separate UT issuing times considered here).

Focusing then on the feature ranking process, the boxplots in Figures 2 and 3 show the top 10 features ordered by their mean RFE ranking, obtained by HLA and RF over the 100 random realizations for each of the four forecast issuing times. Specifically, Figure 2 refers to the prediction of *GOES* C1+ flares, while Figure 3 refers to the prediction of *GOES* M1+ flares.

From these results it becomes possible to assess the impact of feature selection on the prediction performance, by computing specific skill scores and statistics in a cumulative way. The panels in Figures 4 and 5 plot the TSS values obtained by HLA and RF in the case of one specific data set

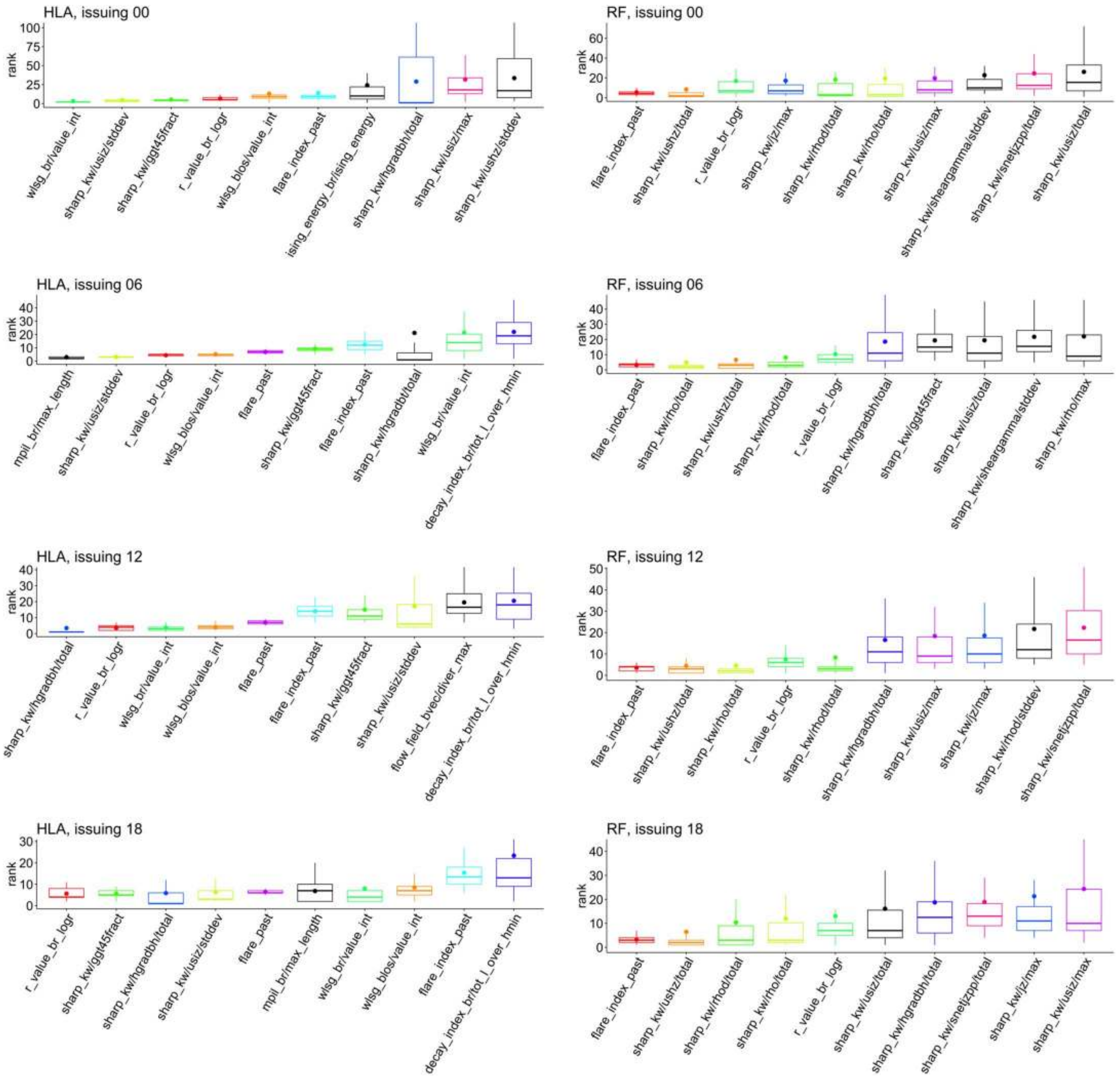

Figure 2. Boxplots of the feature ranks provided by RFE as applied against the outcomes of HLA and RF for the 100 realizations of the training set. The panels show separately the result of the two learning methods (HLA: left column; RF: right column) for the four issuing times considered in the experiment. The focus here is on the prediction of *GOES* C1+ flares.

realization, while adding one feature at a time, starting from the feature with the highest ranking, down to the feature with the tenth highest ranking. A given feature has the same color throughout each set of plots, for all issuing times.

In order to have a clearer picture of the features that repeatedly show the highest predictive impact, the histograms in Figure 6 compute the number of times over the four issuing times that each feature appears in the top-10 ranking of training set averages. These plots only present features that reach the top-10 ranking at least twice out of the four issuing times for a given machine-learning method and flare class. For the C1+ flare prediction (Figure 6; top row) one sees, for example, that the past flare history (*flare_index_past*) and Schrijver's (Schrijver 2007) *R*-value (*r_value_br_logr*) consistently appear for both HLA and RF. This is not the case for the prediction of M1+ flares (Figure 6; bottom row). It should be noted that the importance of the specific features may only be due to the machine-learning method used; it is their consistency of appearance, however, that is notable.

4. Discussion of Results

We first notice that the maximum values of HSS and TSS achieved in Table 1 are distinctly different from one, indicating far from perfect performance. Interestingly, these scores are

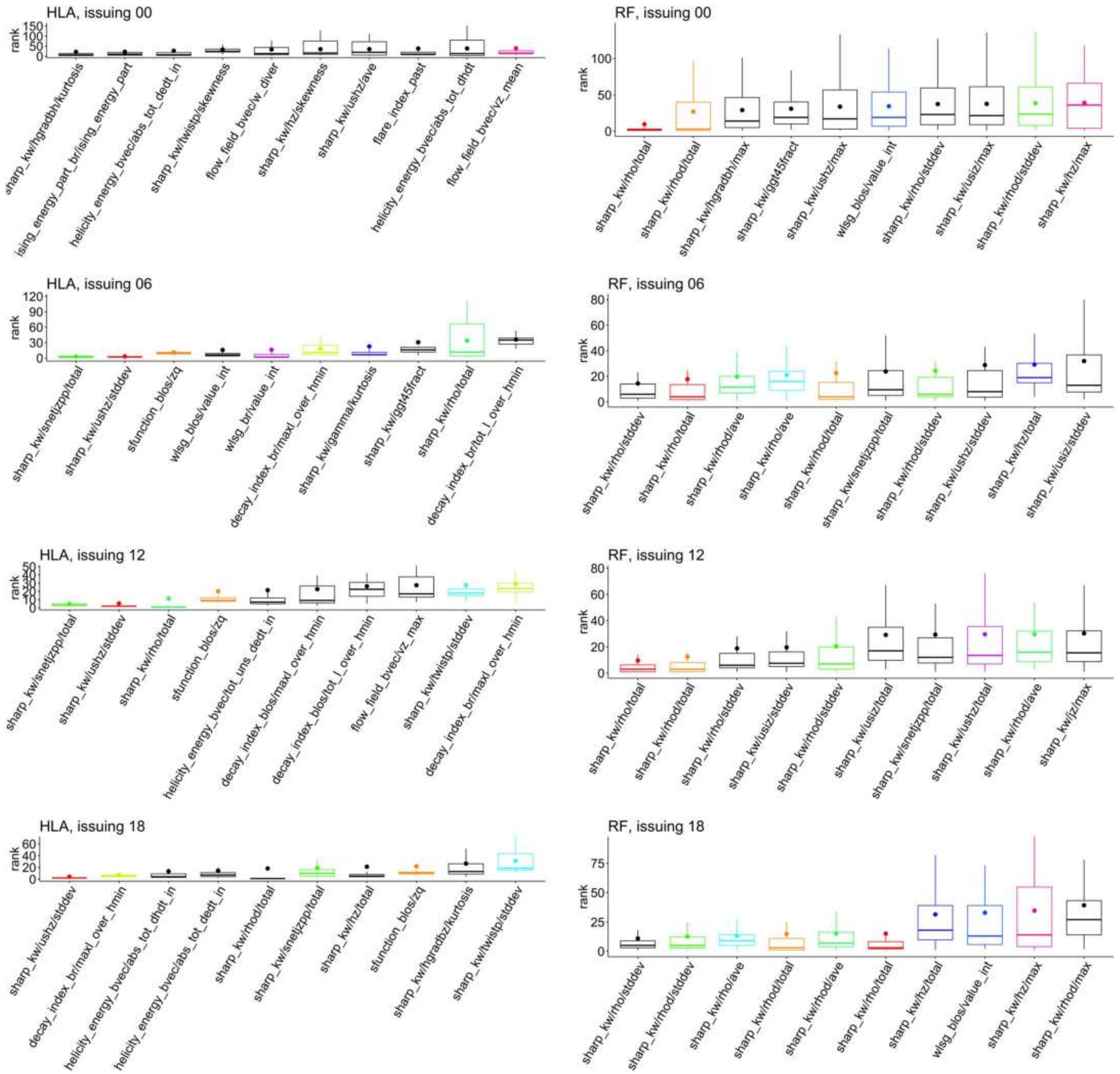

Figure 3. Same as Figure 2, with the focus now being on *GOES* M1+ flares.

almost systematically smaller than the ones recently achieved by methods illustrated in Bobra & Couvidat (2015) and Florios et al. (2018) that use input data with a significantly smaller dimensionality. The methods described in those papers are all supervised, utilize features extracted from HMI data, and perform predictions in a 24 h window. However, the way data preparation is performed and, in particular, how the training set is constructed is significantly different than what is done in the present paper. In particular:

1. The test sets utilized in this work to assess the performances of HLA and RF do not contain feature vectors belonging to ARs with feature vectors contained in the training sets. Instead, the training sets utilized in Bobra & Couvidat (2015), and Florios et al. (2018) combined feature vectors belonging to the same ARs in the two sets.
2. We constructed four separate training/testing sets, each corresponding to a specific UT forecast issuing time on all of the days considered. Our results show reasonably consistent forecast performance across these four issuing-time sets. However, the main benefit to this approach is in the interpretation of the feature selection results. Identifying key forecasting features through their appearance in all (or most) of the top-10 feature ranking lists across these four issuing-time sets increases their robustness through temporal consistency.

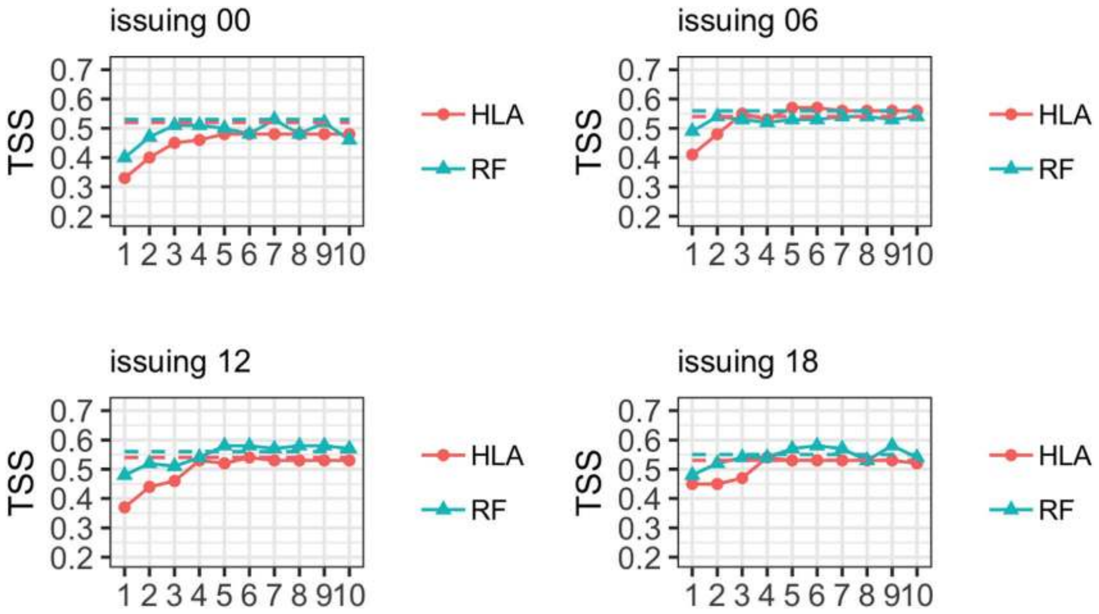

Figure 4. TSS scores obtained by using just the 10 features with the best rankings in decreasing order, from 1 to 10, for both machine-learning methods and all four issuing times, in the case of a specific realization of the test set. Features are added one at a time. The plots refer to the prediction of *GOES* C1+ flares. The dashed horizontal lines are the TSS values obtained by HLA and RF when applied to all 171 features.

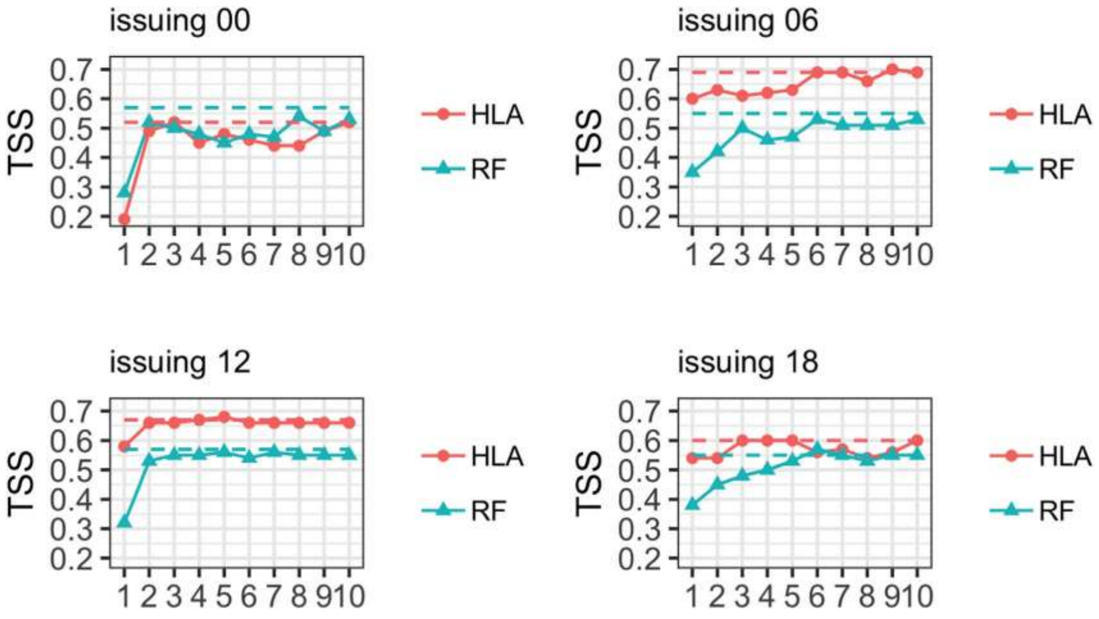

Figure 5. Same as Figure 4, but for *GOES* M1+ flares.

3. The training set utilized in Florios et al. (2018) is populated with approximately the same number of vectors as the test set, while in our approach (and in the one followed in Bobra & Couvidat 2015) the machine-learning methods are trained with training sets two times more populated than the test sets, which is more realistic with respect to typical experimental settings.
4. Our prediction methods are optimized using a fuzzy clustering technique, while in the other three cases the input parameters are fixed in such a way to optimize a specific skill score (namely, the TSS).

In order to assess the impact of these differences against the methods' performance, we trained HLA and RF using all 14,931 point-in-time feature vectors distributing them between training and test sets as was done in Bobra & Couvidat (2015) and Florios et al. (2018). Specifically, we generated the training and test sets focusing on feature vectors instead of on ARs, i.e., we randomly extracted the feature vectors from the database at disposal without imposing any constraint that forbids feature vectors of the same AR to populate both the training and the test set (the two sets are populated as in Bobra & Couvidat 2015, following a 2:1 proportion). Furthermore, we did not care for time consistency and so we mixed up feature vectors belonging to different issuing times. Finally, the prediction methods are optimized in such a way to maximize the TSS. Table 2 shows that TSS increases significantly in the prediction of C1+ and M1+ flares for both

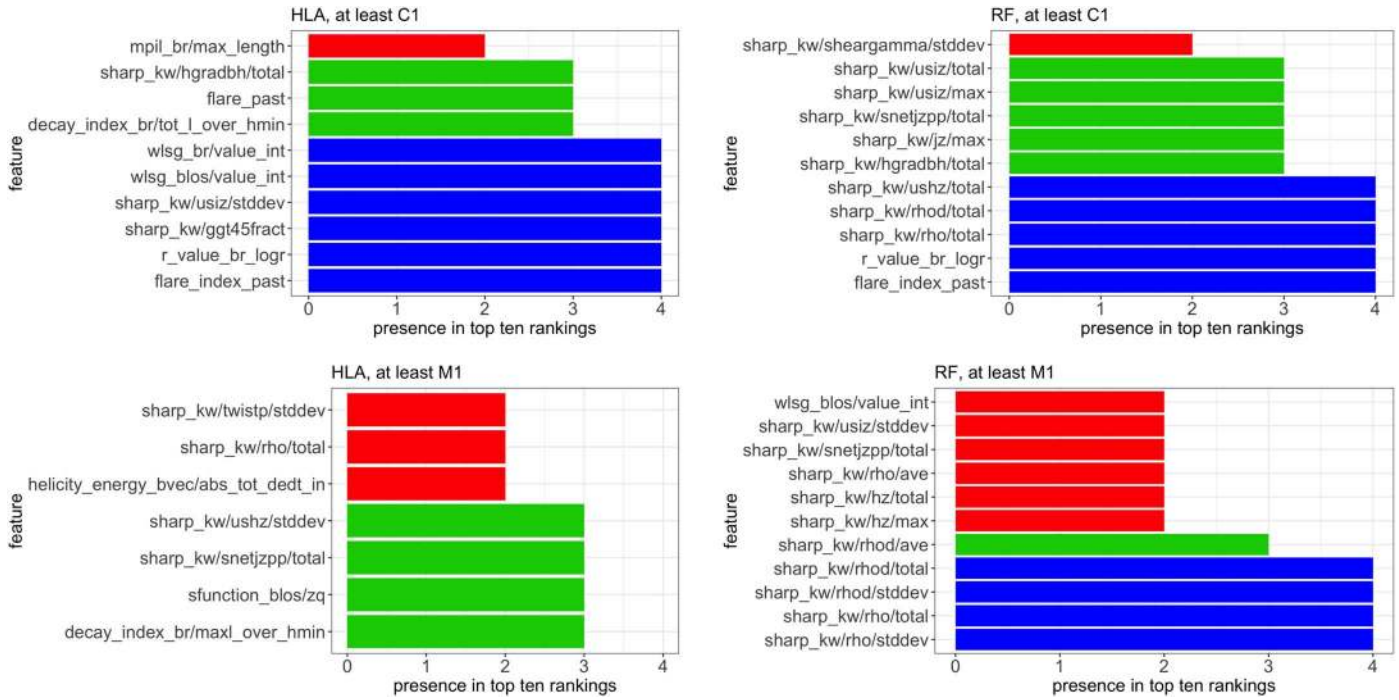

Figure 6. Histograms counting the number of times each feature is selected in the top-10 rankings, on average over the 100 random realizations of the test set, for all issuing times and considering both HLA (left column) and RF (right column) as learning machines. Predictions of *GOES* C1+ flares and *GOES* M1+ flares are shown in the top and bottom rows, respectively.

Table 2
Average TSS- and HSS-values, Along with Applicable Standard Deviations, Over the Outcomes of HLA and RF as Applied Against 100 Random Realizations of the Training/Test Sets

	Test Set-C1+	Test Set C1+	Test Set-M1+	Test Set-M1+
	TSS	HSS	TSS	HSS
HLA	0.58 ± 0.01	0.51 ± 0.01	0.70 ± 0.02	0.31 ± 0.03
RF	0.61 ± 0.01	0.56 ± 0.02	0.71 ± 0.03	0.39 ± 0.02
Florios et al. (2018)	0.60 ± 0.01	0.59 ± 0.01	0.74 ± 0.02	0.49 ± 0.01
Bobra & Couvidat (2015)	⋯	⋯	0.76 ± 0.04	0.52 ± 0.04

Note. The training sets have been generated according to the same procedure as in Bobra & Couvidat (2015) and Florios et al. (2018). The scores presented in those papers are reported in this table.

HLA and RF and also HSS produced by RF becomes larger, although less significantly. These scores are now more in line with the ones obtained in the other two papers, at the same time showing smaller standard deviations. This leads to the conclusion that, not surprisingly, also in flare prediction the biases introduced in the process of training set generation and the way the algorithms are optimized strongly influence the performance of supervised methods.

As far as feature ranking is concerned, it is evident from Figures 2 and 3 that first, features with the best ranking have the smallest standard deviations, so their impact on prediction is consistently high, regardless of the splitting of the data used for training the machine-learning algorithms. In the case of prediction of *GOES* C1+ flares, colors largely repeat in all four panels, telling us that features with highest predictive power are common to all issuing times considered. This behavior is not as robust in the case of prediction of *GOES* M1+ flares, but we are confident that this is a consequence of the lower occurrence rate of M1+ flares and the resulting variation in the flare /no-flare imbalance ratio of the random training sets. A consistent imbalance ratio is more or less guaranteed for C1+ flares, whose comprehensive statistics over solar cycle 24 ensure a well-balanced training process.

Figures 4 and 5 show that only a small number of features (up to 10) over the scores of features proposed and/or applied for flare prediction, are sufficient to achieve maximum performance of a given machine-learning method. Notice from these figures that the highest-ranking feature alone (feature 1) suffices to give TSS and HSS values that are at least half of the maxima achieved. Up to the fourth feature, the values of TSS and HSS saturate already, indicating that adding more features will not improve (and may, in fact, be detrimental to) prediction performance. Also, provided that flare statistics are sufficient to deal with the flare/no-flare imbalance ratio in the random selection of training and test sets, these few best-performing features are consistent for a given prediction method. In their study, Bobra & Couvidat (2015) found that the four most significant features in their analysis were the total unsigned current helicity, total magnitude of the Lorentz force, total photospheric magnetic free energy density, and unsigned vertical current. In our study, Figure 6 shows that features associated with the unsigned vertical current (i.e., total, maximum, standard deviation) are among the most temporally consistent of our best-performing features, particularly for C1+ flares (its standard deviation is in the top-ten features of all four UT issuing times for Hybrid LASSO, while its total and maximum are in the top-10 of three of the four UT issuing times for RF) and less so for M1+ flares (its standard deviation is in the top-10 for two of the four UT issuing times). However, the same figure shows that these predictors may well change from method to method, which hampers efforts to understand physically why some features work better than others. Best performers appear to change also for the prediction of different flare classes, which is a very interesting finding that undoubtedly warrants additional investigation in the future.

5. Conclusions

This study employs the highest-dimension data set of prediction features to date in regards to solar flare forecasting, while it shows TSS values similar to the better performing region-by-region forecasting systems in the literature. This point taken, the actual HSS and TSS values are not identical to —and may even be somewhat lower than—the respective values reported in Bobra & Couvidat (2015) and in Florios et al. (2018), with the latter study using RF applied to FLARECAST data, as well. The reason is that training and testing of machine-learning methods in the present study were not only performed on nonoverlapping data as in previous studies, but even the solar ARs selected for training and testing were different. This conclusion seems in line with the considerations contained in Camporeale (2019), whose careful assessment of the causality issue identifies this as one of the crucial aspects impacting the forecasting performances.

The rationale for using hybrid LASSO and RF in this work is these methods' ability to perform feature ranking via RFE, among other methods (i.e., Fisher's score, Gini index, etc.). However, there are 26 machine-learning methods implemented in FLARECAST. Their definitive evaluation is in progress, so the values of pertinent skill scores may well increase in future studies utilizing FLARECAST data, in the search for finding the optimal machine-learning method(s) for the near-realtime FLARECAST forecasting service. We also understand that a meaningful methodological next step would be to introduce deep learning methods in the pipeline. However, interestingly, the use of these more modern approaches in flare forecasting does not necessarily imply significantly higher skill scores (see, e.g., (Nishizuka et al. 2018), where TSS and HSS for the prediction of M1+ flares are reported as 0.80 and 0.26, respectively).

What will, most likely, not change in the foreseeable future are the following two core conclusions of this work.

First, the current range of properties that have been extracted from the HMI magnetograms show significant redundancy and no more than 10 features contained in these properties are sufficient to allow machine-learning methods to achieve maximum performance.

Second, and perhaps foremost, the maximum values of HSS and TSS achieved are distinctly different from one, indicating far from perfect performance. In physical terms, even using the largest flare prediction data volume assembled to date, we have not managed to substantially surpass the performance of a random-chance forecast (as shown by HSS) or to substantially increase the probability of detection (0.57–0.65 for C1+) despite an encouragingly low probability of false detection (∼0.10). The latter two compete against each other to result in TSS. This is equivalent to saying that flare prediction remains probabilistic, rather than binary yes/no with a perfect performance. The core reasons for this may be multiple: first, we only rely on photospheric magnetic field data, but flares occur above the line-tied photosphere in the low solar corona. Second, flares may well be intrinsically stochastic phenomena, as adopted in a long-standing working hypothesis (Rosner & Vaiana 1978), shown conclusively by the flares' time-dependent Poisson waiting times (Crosby et al. 1998; Wheatland & Litvinenko 2002) and interpreted physically via the concept of self-organized criticality (Lu & Hamilton 1991; Lu et al. 1993; Vlahos et al. 1995)—see also Aschwanden et al. (2016) for a comprehensive review.

This study was enabled by the European Union's Horizon 2020 Research and Innovation Action Flare Likelihood And Region Eruption foreCASTing (FLARECAST) project, under grant agreement No. 640216. All authors warmly acknowledge the support of the FLARECAST project.

Appendix

Here we provide details of the FLARECAST feature labels used in this work, with short descriptions and references to their original definition/implementation (or, e.g., detection methods used in their calculation). Features are grouped in the following manner: Table 3 contains those features derived from $B_{\rm los}$ only and $B_{\rm radial}$ only (Georgoulis 2005, 2012; Georgoulis & Rust 2007; Schrijver 2007; Falconer et al. 2008; Hewett et al. 2008; Ahmed et al. 2010; Mason & Hoeksema 2010; Georgoulis et al. 2012; Zuccarello et al. 2014; Guerra et al. 2015; Kontogiannis et al. 2018); Table 4 contains those features requiring all three vectormagnetic field components (Kusano et al. 2002; Schuck 2008); Table 5 contains only those features related to the total and mean quantities provided as the SHARP keywords of Bobra et al. (2014).

Table 3
FLARECAST *B*_los- and *B*_*r*-derived Feature List with Short Descriptions

Feature Label	Description
alpha_exp_fft_blos/alpha, alpha_exp_fft_br/alpha	Fourier power spectral index
alpha_exp_cwt_blos/alpha, alpha_exp_cwt_br/alpha	Continuous wavelet transform power spectral index
beff_blos/beff, beff_br/beff	Effective connected magnetic field strength $B_{\rm eff}$
decay_index_blos/max_l_over_hmin decay_index_br/max_l_over_hmin	Max. ratio of MPIL length to min. height of critical decay index $l/h(n_{\rm cr})_{\rm min}$
decay_index_blos/tot_l_over_hmin, decay_index_br/tot_l_over_hmin	Total of all separate MPIL ratios of $l/h(n_{\rm cr})_{\rm min}$
decay_index_blos/l_over_minhmin, decay_index_br/l_over_minhmin	Ratio of MPIL $l/h(n_{\rm cr})_{\rm min}$ (for MPIL having lowest $h(n_{\rm cr})_{\rm min}$)
decay_index_blos/maxl_over_hmin, decay_index_br/maxl_over_hmin	Ratio of MPIL $l/h(n_{\rm cr})_{\rm min}$ (for longest MPIL)
flare_past	Binary flag for occurrence of $\geqslant$1 flare in previous 24 hr
flare_index_past	Accumulated *GOES* flare peak magnitudes in previous 24 hr
frdim_blos/frdim, frdim_br/frdim	Fractal dimension
gs_slf/g_s	Sum of the horizontal magnetic gradient
gs_slf/slf	Separation distance lead. and follow. polarity subgroups
ising_energy_blos/ising_energy, ising_energy_br/ising_energy	Ising energy (calculated pixel-by-pixel)
ising_energy_part_blos/ising_energy_part, ising_energy_part_br/ising_energy_part	Ising energy (calculated using $B_{\rm eff}$ flux partitions)
lat_hg	Heliographic latitude of SHARP centroid
lon_hg	Heliographic longitude of SHARP centroid
mf_spectrum_blos/dq, mf_spectrum_br/dq	Multifractal generalized correlation dimension spectrum
mpil_blos/max_length, mpil_br/max_length	Maximum length of a single MPIL
mpil_blos/tot_length, mpil_br/tot_length	Total length of all MPILs
mpil_blos/tot_usflux, mpil_br/tot_usflux	Total unsigned flux around all MPILs
r_value_blos_logr, r_value_br_logr	Schrijver's *R* ($\log_{10}$ form)
sfunction_blos/zq, sfunction_br/zq	Multifractal structure function inertial range index
wlsg_blos/value_int, wlsg_br/value_int	Falconer's ($^{\rm L}{\rm WL}_{\rm SG}$

Table 4
FLARECAST $B_{r,\theta,\phi}$-derived Feature List with Short Descriptions

Feature Label	Description
flow_field_bvec/diver, flow_field_bvec/diver_max, flow_field_bvec/diver_mean	Flow field divergence (total, maximum, mean)
flow_field_bvec/shear, flow_field_bvec/shear_max, flow_field_bvec/shear_mean	Flow field shear (total, maximum, mean)
flow_field_bvec/v_ mean, flow_field_bvec/v_median	Flow field total velocity magnitude (mean, median)
flow_field_bvec/vz_max, flow_field_bvec/vz_mean	Flow field vertical velocity magnitude (mean, median)
flow_field_bvec/w_diver, flow_field_bvec/w_diver_max, flow_field_bvec/w_diver_mean	Flux-weighted flow field divergence (total, maximum, mean)
flow_field_bvec/w_shear, flow_field_bvec/w_shear_max, flow_field_bvec/w_shear_mean	Flux-weighted flow field shear (total, maximum, mean)
helicity_energy_bvec/abs_tot_dedt	Abs. val. net vertical Poynting flux
helicity_energy_bvec/abs_tot_dedt_in	Abs. val. net vertical Poynting flux (emerg. comp.)
helicity_energy_bvec/abs_tot_dedt_sh	Abs. val. net vertical Poynting flux (shear. comp.)
helicity_energy_bvec/abs_tot_dedt_in_plus_sh	Emerg. + shear. abs. values net vertical Poynting flux
helicity_energy_bvec/abs_tot_dhdt	Abs. val. net vertical helicity flux
helicity_energy_bvec/abs_tot_dhdt_in	Abs. val. net vertical helicity flux (emerg. comp.)
helicity_energy_bvec/abs_tot_dhdt_sh	Abs. val. net vertical helicity flux (emerg. comp.)
helicity_energy_bvec/abs_tot_dhdt_in_plus_sh	Emerg. + shear. abs. values net vertical helicity flux
helicity_energy_bvec/tot_uns_dedt	Total unsigned vertical Poynting flux
helicity_energy_bvec/tot_uns_dedt_in	Tot. unsign. vertical Poynting flux (emerg. comp.)
helicity_energy_bvec/tot_uns_dedt_sh	Tot. unsign. vertical Poynting flux (shear. comp.)

Table 4
(Continued)

Feature Label	Description
helicity_energy_bvec/tot_uns_dhdt	Tot. unsign. vertical helicity flux
helicity_energy_bvec/tot_uns_dhdt_in	Tot. unsign. vertical helicity flux (emerg. comp.)
helicity_energy_bvec/tot_uns_dhdt_sh	Tot. unsign. vertical helicity flux (shear. comp.)
nn_currents/tot_us_cur	Total unsigned nonneutralized currents

Table 5
FLARECAST SHARP-keyword Related Feature List with Short Descriptions

Feature Label	Description
sharp_kw/gamma/ave, sharp_kw/gamma/stddev	Field inclin. ang. (mean, st. dev.)
sharp_kw/gamma/skewness, sharp_kw/gamma/kurtosis	Field inclin. ang. (skewn., kurt.)
sharp_kw/gamma/total, sharp_kw/gamma/max, sharp_kw/gamma/median	Field inclin. ang. (tot., max., med.)
sharp_kw/ggt45fract	% tot. area with shear angle >45°
sharp_kw/hgradbh/ave, sharp_kw/hgradbh/stddev	Horiz. grad. $B_{\rm hor}$ (mean, st. dev.)
sharp_kw/hgradbh/skewness, sharp_kw/hgradbh/kurtosis	Horiz. grad. $B_{\rm hor}$ (skewn., kurt.)
sharp_kw/hgradbh/total, sharp_kw/hgradbh/max, sharp_kw/hgradbh/median	Horiz. grad. $B_{\rm hor}$ (tot, max, med)
sharp_kw/hgradbt/ave, sharp_kw/hgradbt/stddev	Horiz. grad. $B_{\rm tot}$ (mean, st. dev.)
sharp_kw/hgradbt/skewness, sharp_kw/hgradbt/kurtosis	Horiz. grad. $B_{\rm tot}$ (skewn., kurt.)
sharp_kw/hgradbt/total, sharp_kw/hgradbt/max, sharp_kw/hgradbt/median	Horiz. grad. $B_{\rm tot}$ (tot., max., med.)
sharp_kw/hgradbz/ave, sharp_kw/hgradbz/stddev	Horiz. grad. B_r (mean, st. dev.)
sharp_kw/hgradbz/skewness, sharp_kw/hgradbz/kurtosis	Horiz. grad. B_r (skewn., kurt.)
sharp_kw/hgradbz/total, sharp_kw/hgradbz/max, sharp_kw/hgradbz/median	Horiz. grad. B_r (tot., max., med.)
sharp_kw/hz/ave, sharp_kw/hz/stddev	Vert. curr. hel. (mean, st. dev.)
sharp_kw/hz/skewness, sharp_kw/hz/kurtosis	Vert. curr. hel. (skewn., kurt.)
sharp_kw/hz/total, sharp_kw/hz/max, sharp_kw/hz/median	Vert. curr. hel. (tot., max., med.)
sharp_kw/jz/ave, sharp_kw/jz/stddev	Vert. curr. (mean, st. dev.)
sharp_kw/jz/skewness, sharp_kw/jz/kurtosis	Vert. curr. (skewn., kurt.)
sharp_kw/jz/total, sharp_kw/jz/max, sharp_kw/jz/median	Vert. curr. (tot., max., med.)
sharp_kw/rho/ave, sharp_kw/rho/stddev	Photosph. excess magn. en. (mean, st. dev.)
sharp_kw/rho/skewness, sharp_kw/rho/kurtosis	Photosph. excess magn. en. (skewn., kurt.)
sharp_kw/rho/total, sharp_kw/rho/max, sharp_kw/rho/median	Photosph. excess magn. en. (tot., max., med.)
sharp_kw/rhod/ave, sharp_kw/rhod/stddev	Photosph. excess magn. en. dens. (mean, st. dev.)
sharp_kw/rhod/skewness, sharp_kw/rhod/kurtosis	Photosph. excess magn. en. dens. (skewn., kurt.)
sharp_kw/rhod/total, sharp_kw/rhod/max, sharp_kw/rhod/median	Photosph. excess magn. en. dens (tot., max., med.)
sharp_kw/sflux/ave, sharp_kw/sflux/stddev	Signed flux (mean, st. dev.)
sharp_kw/sflux/skewness, sharp_kw/sflux/kurtosis	Signed flux (skewn., kurt.)
sharp_kw/sflux/total, sharp_kw/sflux/max, sharp_kw/sflux/median	Signed flux (tot., max., med.)
sharp_kw/sheargamma/ave, sharp_kw/sheargamma/stddev	$B_{\rm tot}$ shear angle (mean, st. dev.)
sharp_kw/sheargamma/skewness, sharp_kw/sheargamma/kurtosis	$B_{\rm tot}$ shear angle (skewn., kurt.)
sharp_kw/sheargamma/total, sharp_kw/sheargamma/max, sharp_kw/sheargamma/median	$B_{\rm tot}$ shear angle (tot., max., med.)
sharp_kw/snetjzpp/total	Sum abs. val. net currents per polarity
sharp_kw/twistp/ave, sharp_kw/twistp/stddev, sharp_kw/twistp/skewness, sharp_kw/twistp/kurtosis	Twist parameter (mean, st. dev., skewn., kurt.)
sharp_kw/twistp/total, sharp_kw/twistp/max, sharp_kw/twistp/median	Twist parameter (tot., max., med.)
sharp_kw/usflux/ave, sharp_kw/usflux/stddev, sharp_kw/usflux/skewness, sharp_kw/usflux/kurtosis	Uns. flux (mean, st. dev., skewn., kurt.)
sharp_kw/usflux/total, sharp_kw/usflux/max, sharp_kw/usflux/median	Uns. flux (tot., max., med.)
sharp_kw/ushz/ave, sharp_kw/ushz/stddev, sharp_kw/ushz/skewness, sharp_kw/ushz/kurtosis	Uns. vert. curr. hel. (mean, st. dev., skewn., kurt.)
sharp_kw/ushz/total, sharp_kw/ushz/max, sharp_kw/ushz/median	Uns. vert. curr. hel. (tot., max., med.)
sharp_kw/usiz/ave, sharp_kw/usiz/stddev, sharp_kw/usiz/skewness, sharp_kw/usiz/kurtosis	Uns. vert. curr. (mean, st. dev., skewn., kurt.)
sharp_kw/usiz/total, sharp_kw/usiz/max, sharp_kw/usiz/median	Uns. vert. curr. (tot., max., med.)

ORCID iDs

Federico Benvenuto https://orcid.org/0000-0002-4776-0256
D. Shaun Bloomfield https://orcid.org/0000-0002-4183-9895
Manolis K. Georgoulis https://orcid.org/0000-0001-6913-1330

References

Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, SoPh, 283, 157
Ahmed, O. W., Qahwaji, R., Colak, T., Dudok De Wit, T., & Ipson, S. 2010, Vis. Comp., 26, 385
Aschwanden, M. J., Crosby, N. B., Dimitropoulou, M., et al. 2016, SSRv, 198, 47
Barnes, G., & Leka, K. D. 2008, ApJL, 688, L107
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89
Benvenuto, F., Piana, M., Campi, C., & Massone, A. M. 2018, ApJ, 853, 90
Benz, A. O. 2017, LRSP, 14, 2
Bezdek, J. C., Ehrlich, R., & Full, W. 1984, CG, 10, 191
Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJL, 747, L41
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Breiman, L. 2001, Machine Learning, 45, 5
Camporeale, E. 2019, arXiv:1903.05192
Colak, T., & Qahwaji, R. 2009, SpWea, 7, S06001
Cortes, C., & Vapnik, V. 1995, Machine Learning, 20, 273
Crosby, N., Vilmer, N., Lund, N., & Sunyaev, R. 1998, A&A, 334, 299
Evgeniou, T., Pontil, M., & Poggio, T. 2000, Adv. Comput. Math., 13, 1
Falconer, D. A., Moore, R. L., & Gary, G. A. 2008, ApJ, 689, 1433
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, SoPh, 293, 28
Gallagher, P. T., Moon, Y. J., & Wang, H. 2002, SoPh, 209, 171
Georgoulis, M. K. 2005, SoPh, 228, 5
Georgoulis, M. K. 2012, SoPh, 276, 161
Georgoulis, M. K., & Rust, D. M. 2007, ApJL, 661, L109
Georgoulis, M. K., Titov, V. S., & Mikić, Z. 2012, ApJ, 761, 61
Guerra, J. A., Park, S. H., Gallagher, P. T., et al. 2018, SoPh, 293, 9
Guerra, J. A., Pulkkinen, A., Uritsky, V. M., & Yashiro, S. 2015, SoPh, 290, 335
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. 2002, Machine Learning, 46, 389
Hewett, R. J., Gallagher, P. T., McAteer, R. T. J., et al. 2008, SoPh, 248, 311
Higgins, P. A., Gallagher, P. T., McAteer, R. T. J., & Bloomfield, D. S. 2011, AdSpR, 47, 2105
Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
Kontogiannis, I., Georgoulis, M. K., Park, S.-H., & Guerra, J. A. 2018, SoPh, 293, 96
Korsos, M. B., & Erdelyi, R. 2016, ApJ, 823, 153
Kusano, K., Maeshiro, T., Yokoyama, T., & Sakurai, T. 2002, ApJ, 577, 501
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, R., Cui, Y., He, H., & Wang, H. 2008, AdSpR, 42, 1469
Li, R., Wang, H. N., He, H., Cui, Y. M., & Du, Z. L. 2007, ChJAA, 7, 441
Lu, E. T., & Hamilton, R. J. 1991, ApJL, 380, L89
Lu, E. T., Hamilton, R. J., McTiernan, J. M., & Bromund, K. R. 1993, ApJ, 412, 841
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
Massone, A. M., Piana, M. & The FLARECAST Team 2018, in Machine Learning for Flare Forecasting, Machine Learning Techniques for Space Weather, ed. E. Camporeale, S. Wing, & J. R. Johnson (Amsterdam: Elsevier)
McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2017, JSWSC, 8, A34
Murray, S. A., Bingham, S., Sharpe, M., & Jackson, D. R. 2017, SpWea, 15, 577
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. 2018, ApJ, 858, 113
Park, E., Moon, Y. J., Shin, S., et al. 2018, ApJ, 869, 91
Priest, E. R., & Forbes, T. G. 2002, A&ARv, 10, 313
Rosner, R., & Vaiana, G. S. 1978, ApJ, 222, 1104
Sadykov, V. M., & Kosovichev, A. G. 2017, ApJ, 849, 148
Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, SoPh, 275, 207
Schrijver, C. J. 2007, ApJL, 655, L117
Schuck, P. W. 2008, ApJ, 683, 1134
Schwenn, R. 2006, LRSP, 3, 2
Tibshirani, R. 1996, J. Royal. Stat. Soc. B, 58, 267
Vlahos, L., Georgoulis, M., Kluiving, R., & Paschos, P. 1995, A&A, 299, 897
Wang, H. N., Cui, Y. M., Li, R., Zhang, L. Y., & Han, H. 2008, AdSpR, 42, 1464
Wheatland, M. S., & Litvinenko, Y. E. 2002, SoPh, 211, 255
Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., & Lange, K. 2009, Bioinformatics, 25, 714
Youden, W. J. 1950, Cancer, 3, 32
Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, SoPh, 255, 91
Zuccarello, F. P., Seaton, D. B., Mierla, M., et al. 2014, ApJ, 785, 88