A&A 639, A44 (2020)
https://doi.org/10.1051/0004-6361/201937426

Astronomy & Astrophysics

Distinguishing between flaring and nonflaring active regions

Soumitra Hazra[1,2,3], Gopal Sardar[4,5], and Partha Chowdhury[6]

[1] Université Paris-Saclay, CNRS, CEA, Astrophysique, Instrumentation et Modélisation de Paris-Saclay, 91191 Gif-sur-Yvette, France
e-mail: soumitra.hazra@cea.fr, soumitra.hazra@gmail.com
[2] Université Paris-Saclay, CNRS, Institut d'Astrophysique Spatiale, 91405 Orsay, France
[3] Center of Excellence and Space Sciences India, Indian Institute of Science Education and Research Kolkata, Mohanpur 741246, West Bengal, India
[4] Department of Physical Sciences, Indian Institute of Technology Jodhpur, Jodhpur 342011, India
[5] Department of Physical Sciences, Indian Institute of Science Education and Research Kolkata, Mohanpur 741246, West Bengal, India
[6] University College of Science and Technology, Department of Chemical Technology, University of Calcutta, 92, A.P.C. Road, Kolkata 700009, West Bengal, India

Received 28 December 2019 / Accepted 19 May 2020

ABSTRACT

Context. Large-scale solar eruptions significantly affect space weather and damage space-based human infrastructures. It is necessary to predict large-scale solar eruptions; it will enable us to protect the vulnerable infrastructures of our modern society.
Aims. We investigate the difference between flaring and nonflaring active regions. We also investigate whether it is possible to forecast a solar flare.
Methods. We used photospheric vector magnetogram data from the Solar Dynamic Observatory's Helioseismic Magnetic Imager to study the time evolution of photospheric magnetic parameters on the solar surface. We built a database of flaring and nonflaring active regions observed on the solar surface from 2010 to 2017. We trained a machine-learning algorithm with the time evolution of these active region parameters. Finally, we estimated the performance obtained from the machine-learning algorithm.
Results. The strength of some magnetic parameters such as the total unsigned magnetic flux, the total unsigned magnetic helicity, the total unsigned vertical current, and the total photospheric magnetic energy density in flaring active regions are much higher than those of the non-flaring regions. These magnetic parameters in a flaring active region evolve fast and are complex. We are able to obtain a good forecasting capability with a relatively high value of true skill statistic. We also find that time evolution of the total unsigned magnetic helicity and the total unsigned magnetic flux provides a very high ability of distinguishing flaring and nonflaring active regions.
Conclusions. We can distinguish a flaring active region from a nonflaring region with good accuracy. We confirm that there is no single common parameter that can distinguish all flaring active regions from the nonflaring regions. However, the time evolution of the top two magnetic parameters, the total unsigned magnetic flux and the total unsigned magnetic helicity, have a very high distinguishing capability.

Key words. methods: data analysis – methods: observational – Sun: flares – Sun: coronal mass ejections (CMEs) – Sun: magnetic fields

1. Introduction

Solar flares and coronal mass ejections are the two greatest explosions in the Solar System. These two explosions release a huge amount of magnetic energy into the solar corona, creating disturbances in space weather. These two events directly affect the Earth's atmosphere, causing geomagnetic disturbances. It is now well known that the magnetic field structures in the Sun are responsible for large-scale eruptions. The study of magnetic fields in the Sun is critical for understanding the energy buildup and release mechanism in solar flares and coronal mass ejection.

Solar flares and coronal mass ejections are believed to be a storage-and-release mechanism by which the nonpotential magnetic field of the solar corona is released abruptly (Priest & Forbes 2002; Shibata & Magara 2011). It is also believed that complex magnetic structures on the solar surface are related to the onset of solar eruptions. Many studies have been performed to investigate the relationship between solar eruptions and photospheric magnetic parameters. Many active region parameters have been proposed to characterize the nonpotentiality of the magnetic field structures on the solar surface. Some of the well-known nonpotentiality parameters are the current helicity (Abramenko et al. 1996; Zhang & Bao 1999), the vertical electric current (Leka et al. 1993), the horizontal gradient of the longitudinal magnetic field (Zirin & Wang 1993; Tian et al. 2002), the total photospheric magnetic free-energy density (Wang et al. 1996; Metcalf et al. 2005), the strong magnetic shear (Low 1977; Kusano et al. 1995), the reverse magnetic shear (Kusano et al. 2004; Vemareddy et al. 2012), the shear angle (Ambastha et al. 1993), and the twist parameter (Pevtsov et al. 1994; Holder et al. 2004). Although individual case studies indicate a strong relationship between these nonpotentiality parameters and the flare productivity, it is unclear so far which property is common in all the eruptive active regions and distinguishes them from other noneruptive active regions.

It is now well known that magnetic field structures on the solar surface change significantly with time. A detailed study of this photospheric magnetic field evolution may shed light on the energy buildup and release mechanism of solar eruptions. The most frequently discussed mechanism for the change in the photospheric magnetic field structure is flux emergence and cancellation (Livi et al. 1989; Spirock et al. 2002; Sudol & Harvey 2005; Burtseva & Petrie 2013). Flux emergence and cancellation are found to play a significant role in some theories of solar eruptions (van Ballegooijen & Martens 1989; Amari et al. 2010). Flux cancellation is also one of the necessary conditions for the formation of solar filaments (Martin et al. 1985; Gaizauskas et al. 1997; Martens & Zwaan 2001). Solar filaments are believed to be one of the main precursors for solar eruptions (Sinha et al. 2019). The possible orientation of the magnetic field that is ejected as a result of solar eruptions can be predicted by studying the hemispheric preference of the filament chirality (Martin et al. 1994; Hazra et al. 2018). In summary, the time evolution of the photospheric magnetic field parameters plays an important role in the onset phase of the solar flare. However, because of the large amount of solar data, it is almost impossible to analyze every solar eruptive event manually. We must build some reliable automated method that can analyze the eruptive active regions and distinguish them from other noneruptive active regions.

In recent times, machine-learning appears as a promising automated candidate to reliably forecast solar eruptive events (Ahmed et al. 2013; Bobra & Couvidat 2015; Bobra & Ilonidis 2016; Nishizuka et al. 2017; Hamdi et al. 2017; Ma et al. 2017; Filali Boubrahimi & Angryk 2018; Florios et al. 2018; Inceoglu et al. 2018). Machine-learning is also used to identify the common parameter that is most important to distinguish an eruptive active region from other noneruptive active regions. Dhuri et al. (2019) used machine-learning to determine the critical criteria in the onset phase that can lead to a solar flare. Different types of data sets are used for the purpose of predicting eruptive events using machine-learning. Yu et al. (2009) and Yuan et al. (2010) used line-of-sight magnetogram data obtained from Michelson Doppler Imager (MDI) for flare prediction. Aggarwal et al. (2018) used filament metadata for the prediction of eruptive events. However, most of the studies used vector magnetogram data obtained from the Helioseismic Magnetic Imager (HMI) onboard the Solar Dynamics Observatory (SDO) for the purpose of flare prediction. Different machine-learning classifiers have been used for the solar flare prediction. Some studies also used the time series of the magnetic field data obtained from HMI for flare prediction (Hamdi et al. 2017; Ma et al. 2017).

In this paper, we aim to investigate the importance of the time evolution of magnetic parameters in terms of flare forecasting. We find that there is a significant difference between an eruptive and a noneruptive active region in terms of both strength and time evolution of photospheric magnetic parameters. We also determine the common magnetic parameter that will clearly separate an eruptive active regions from a noneruptive regions. For this purpose, we trained a machine-learning algorithm using the time evolution of the photospheric magnetic parameters. We are able to predict a solar flare quite well. We find that total unsigned magnetic helicity and total unsigned magnetic flux have a higher distinguishing capability than other photospheric magnetic parameters.

Section 2 describes the details of the data we used. We present a detailed manual comparison study between eruptive and noneruptive active regions in terms of the time evolution of magnetic parameters in Sect. 3. We present a comparison study between eruptive and noneruptive active regions using the machine-learning algorithm in Sect. 4. We also describe the details of the machine-learning algorithm and its performance in Sect. 4. Finally, we present a summary and our conclusions in Sect. 5.

2. Data

2.1. Data for active regions

The HMI, an instrument on board the SDO spacecraft, provides us continuous full-disk photospheric magnetic field data (Scherrer et al. 2012; Schou et al. 2012). The HMI team developed an automated method that detects active region patches from the full-disk vector magnetogram data and provides us derivative data that are called space-weather HMI active region patches (SHARP) (Bobra et al. 2014). The automatic-detection algorithm operates on the line-of-sight magnetic field image and creates a smooth bounding curve, called bitmap, which is centered on the flux-weighted centroid. The HMI Stokes I, Q, U, V data were inverted within the smooth bounding curve with the code called very fast inversion of the Stokes vector (VFISV), which is based on the Middle-Eddington model of the solar atmosphere. The 180° ambiguity in the transverse component of the magnetic field was corrected for using the minimum-energy algorithm (Metcalf 1994; Crouch et al. 2009). The inverted and disambiguated magnetic vector field data were remapped to a Lambert cylindrical equal-area projection, which gives us decomposed Bx, By, and Bz data. JSOC provides us these decomposed data. We downloaded these decomposed data from the JSOC webpage. We calculated 17 active region magnetic field parameters every 12 min from these SHARP data. These parameters are listed with keywords and formula in Table 1. We followed the same procedure to calculate the active region magnetic field parameters as defined in Bobra & Couvidat (2015). We considered the pixels that are within bitmap and above a high-confidence disambiguation threshold (coded value is greater than 60) for our magnetic parameter calculation. We used a finite-difference method to calculate the computational derivative needed for the parameter calculation. We used Green's function technique with a monopole depth of 0.00001 pixels to calculate the potential magnetic field, which is necessary for the calculation of the total photospheric magnetic free-energy density. We neglected active regions near the limb, where it is difficult to see magnetic structures because of the projection effect. The calculated magnetic field parameter data are also not reliable near the limb. We therefore only considered the data for our study that were within ±70° from the disk center. We note that data for all of these magnetic parameters are available in the SHARP header[1].

2.2. Data for a solar flare

We considered the solar flare for our study based on the peak X-ray flux observed by GOES X-ray satellites. When the GOES satellite detects a flare, it generally reports this to the flare catalog. Then the flare is paired with its parent active region. Generally, five types of flares namely A, B, C, M, and X are observed by GOES satellites. While X and M class are high-intensity flares (intensity greater than 10^{-5} Wm^{-2}); other A, B, C flares

[1] SHARP data products from SDO HMI can be found at jsoc.stanford.edu (Scherrer et al. 2012).

Table 1. Keywords, definitions and units of our calculated active region parameters.

Keyword	Description	Unit	Formula
TOTUSJH	Total unsigned current helicity	$G^2\,m^{-1}$	$H_{c_{\rm total}} \propto \sum \lvert B_z.J_z\rvert$
TOTPOT	Total photospheric magnetic free-energy density	$ergs\,cm^{-1}$	$\rho_{\rm tot} \propto \sum \left(\boldsymbol{B}^{\rm obs} - \boldsymbol{B}^{\rm pot}\right)^2 dA$
TOTUSJZ	Total unsigned vertical current	Amperes	$J_{z,\rm total} \propto \sum \lvert J_z\rvert dA$
SVANCPP	Sum of the modulus of the net current per polarity	Amperes	$J_{z,\rm sum} \propto \lvert \sum^{B_z^+} J_z dA\rvert + \lvert \sum^{B_z^-} J_z dA\rvert$
ABSNZH	Absolute value of the net current helicity	$G^2\,m^{-1}$	$H_{c_{\rm abs}} \propto \lvert \sum B_z.J_z\rvert$
USFLUX	Total unsigned flux	Maxwell	$\Phi \propto \sum \lvert B_z\rvert dA$
MEANPOT	Mean photospheric magnetic free-energy density	$ergs\,cm^{-3}$	$\bar{\rho} \propto \frac{1}{N}\sum \left(\boldsymbol{B}^{\rm obs} - \boldsymbol{B}^{\rm pot}\right)^2$
MEANGAM	Mean angle of the field from radial	Degrees	$\bar{\gamma} \propto \frac{1}{N}\sum \arctan\left(\frac{B_{\rm h}}{B_z}\right)$
MEANSHR	Mean shear angle	Degrees	$\bar{\Gamma} \propto \frac{1}{N}\sum \arccos\left(\frac{\boldsymbol{B}^{\rm obs}.\boldsymbol{B}^{\rm pot}}{\lvert B^{\rm obs}\rvert.\lvert B^{\rm pot}\rvert}\right)$
SHRGT45	Fraction of area with shear >45°	m^2	Area with shear > 45°/total area
AREA_ACR	Area of strong-field pixels in an active region	m^2	Area = ∑Pixels
MEANGBT	Mean gradient of the total field	$G\,Mm^{-1}$	$\overline{\lvert\nabla B_{\rm tot}\rvert} = \frac{1}{N}\sum \sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}$
MEANGBZ	Mean gradient of the vertical field	$G\,Mm^{-1}$	$\overline{\lvert\nabla B_z\rvert} = \frac{1}{N}\sum \sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}$
MEANGBH	Mean gradient of the horizontal field	$G\,Mm^{-1}$	$\overline{\lvert\nabla B_{\rm h}\rvert} = \frac{1}{N}\sum \sqrt{\left(\frac{\partial B_{\rm h}}{\partial x}\right)^2 + \left(\frac{\partial B_{\rm h}}{\partial y}\right)^2}$
MEANJZH	Mean current helicity	$G^2\,m^{-1}$	$\bar{H}_{\rm c} \propto \frac{1}{N}\sum B_z J_z$
MEANJZD	Mean vertical current density	$mA\,m^{-2}$	$J_z \propto \frac{1}{N}\sum \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right)$
MEANALP	Mean characteristic twist parameter, α	1 Mm	$\alpha_{\rm total} \propto \frac{\sum J_z B_z}{\sum B_z^2}$

are less intensive ones. For our study, we only consider X and M class flares as a flare. We also only consider the flares for our study which are within ±70° of the central meridian and if there is also an associated parent active region.

3. Comparing flaring and nonflaring regions

Active regions NOAA 11166 (SHARP 401), NOAA 11283 (SHARP 833), and NOAA 11143 (SHARP 335) were chosen for the comparison study between flaring and nonflaring active regions. All of these active regions transited across the visible solar disk for a long time. AR 11166 produced one X-class and two M-class solar flares during the passage across the visible solar disk, while AR 11283 produced two X-class and five M-class solar flares. In contrast, AR 1143 produced no flare during its transit. The questions now are why these three active regions behaved so differently during their transit across the solar disk, and whether it is possible to distinguish flaring and nonflaring active regions.

It is now well known that the different magnetic nature of the active regions is responsible for different behaviors. Here, we study the temporal evolution of the magnetic parameters of the photospheric active region to form an idea about the difference between flaring and nonflaring active regions. Figures 1–3 show the temporal evolution of four magnetic parameters: the total unsigned magnetic flux ($\Phi_{\rm tot}$), the total unsigned current helicity ($h_{c,\rm tot}$), the total unsigned vertical current ($J_{z,\rm tot}$), and the proxy of the total photospheric magnetic free-energy density ($\rho_{\rm tot}$). All these four parameters have a much higher value in the case of the flaring active regions (AR 11166 and 11283) than for the nonflaring active region (AR 11143). All four magnetic parameters also show significant evolution. The total unsigned magnetic flux for AR 11166 decreases before the first large-scale flare, but increases for the other two flares (Fig. 1). The other three magnetic parameters for AR 11166 show an increasing trend before the first flare. The total unsigned magnetic flux for AR 11283 shows a significant decreasing trend before the first flare and an increasing trend later (Fig. 2). The total unsigned magnetic helicity for the two flaring active regions (AR 11166 and 11283) shows an increasing trend before the first flare, and both active regions start the flaring activity when the value of the magnetic helicity is sufficiently high. Another interesting point is that when an active region starts flaring, it continues to flare for some time. All four magnetic parameters also show significant evolution for nonflaring AR 11283, but have a much lower value than the other two flaring active regions (see Fig. 3). We also note that the signal-to-noise ratio in the data near the solar limb is high, therefore the values of the magnetic parameters in our time evolution study are not reliable for the start and end times (active regions are near the limb).

The change in the total unsigned magnetic flux during the active region transit is mostly due to flux cancellation and emergence on the solar surface. The magnetic flux is always observed to disappear when the magnetic flux of one particular polarity encounters flux fragments of opposite polarity. Some previous studies have indicated that flux cancellation plays an important role in triggering solar eruptions (van Ballegooijen & Martens 1989; Amari et al. 2010). The total unsigned magnetic flux of the active region and the magnetic flux near the polarity-inversion line (R-value) is also found to be correlated well with the flaring activity and the coronal X-ray luminosity (Schrijver 2007; Leka & Barnes 2007; Barnes & Leka 2008; Burtseva & Petrie 2013; Hazra et al. 2015). The emergence of the new magnetic flux is also a well-observed phenomenon and believed to be one of the mechanisms for the formation of the current sheet (Tur & Priest 1976; Wang & Tang 1993; Sudol & Harvey 2005).

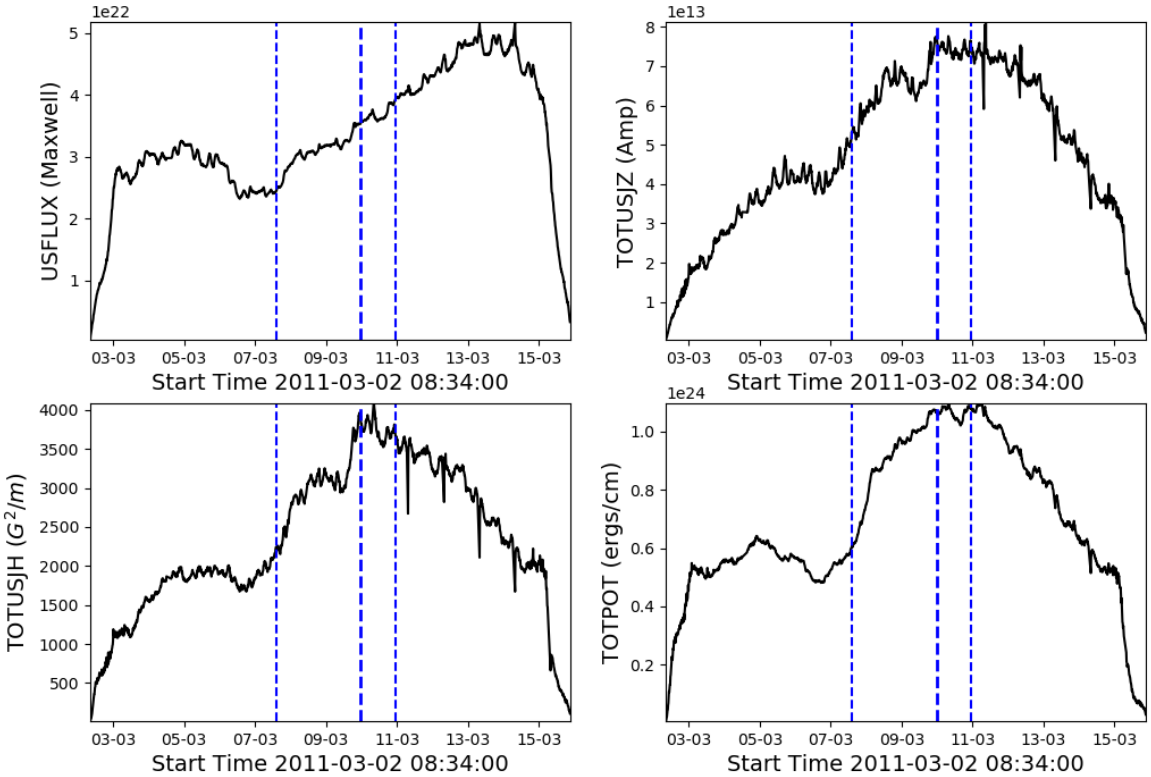

Fig. 1. Time evolution of the total unsigned magnetic flux (USFLUX), the total unsigned vertical current (TOTUSJZ), the total unsigned current helicity (TOTUSJH), and the total photospheric magnetic free-energy density (TOTPOT) for active region NOAA 11166 (SHARP 401). The thick dashed vertical blue line corresponds to an X-class solar flare, and the thin dashed vertical blue line corresponds to an M-class solar flare.

Our results indicate that the total unsigned magnetic flux is considerably higher in flaring active regions than in nonflaring active regions.

We also found that flaring active regions are magnetically more complex than nonflaring active regions. The magnetic complexity in an active region can be characterized by different magnetic parameters: the vertical electric current, magnetic helicity, twist, shear angle, photospheric magnetic free energy density, etc. (Abramenko et al. 1996; Metcalf et al. 2005; Pevtsov et al. 1994; Park et al. 2008). Magnetic helicity, which is a measure of twist, shear, and the inter-linkage of magnetic field lines, is a conserved quantity in an ideal MHD scenario (Berger & Field 1984). A change in magnetic helicity reflects a deviation from the ideal MHD scenario and indicates that the magnetic complexity inside the active regions evolves. We find significantly higher magnetic helicity and excess magnetic free energy in the flaring active region than in the nonflaring region. Recent theoretical and observational studies also suggest that the injection of magnetic helicity of both the same and opposite sign into the global helicity of a system can trigger a solar flare (Kusano et al. 2002; Moon et al. 2002; Park et al. 2008, 2012, 2013). Kusano et al. (2003) also developed a theoretical model of a solar flare based on the destruction of magnetic helicity. Our result indicates both the accumulation and destruction of magnetic helicity before the onset of a solar flare.

In summary, our result highlights that the time evolution of magnetic parameters is important for distinguishing flaring and nonflaring active regions. However, the question arises which magnetic parameter is more important. This is very difficult to say. In reality, a large number of active regions appears on the solar surface within a few days. It is difficult to analyze all active regions manually to predict the probability that on such region erupts. We have to develop some automated method that will help us to predict whether an active region will flare.

4. Comparing flaring and nonflaring regions using machine learning

In the previous section, we discussed the differences between flaring and nonflaring active regions based on the time evolution of magnetic parameters. In this section, we distinguish them based on an automated machine-learning method. Machine learning is a branch of artificial intelligence that provides a computer the ability to learn automatically and improve from past experience without being explicitly programmed. Two types of learning, unsupervised learning and supervised learning, exist in the machine-learning literature. In the unsupervised learning scenario, the machine-learning algorithm identifies the patterns in the data without the use of explicitly labeled information, while in the supervised learning, labeled data are available. As we already have a well-known database of flaring and nonflaring active regions, we used supervised learning techniques for our problem.

4.1. Data preparation

To train the supervised machine-learning algorithms, we first defined the positive and negative class. We followed the definition prescribed by Ma et al. (2017) and Dhuri et al. (2019) for this purpose. An active region that produces at least one X- or M-class solar flare during its transit across the visible solar disk belongs to the positive class, and an active region that does not produce any X- or M-class flare during its transit belongs to the negative class.

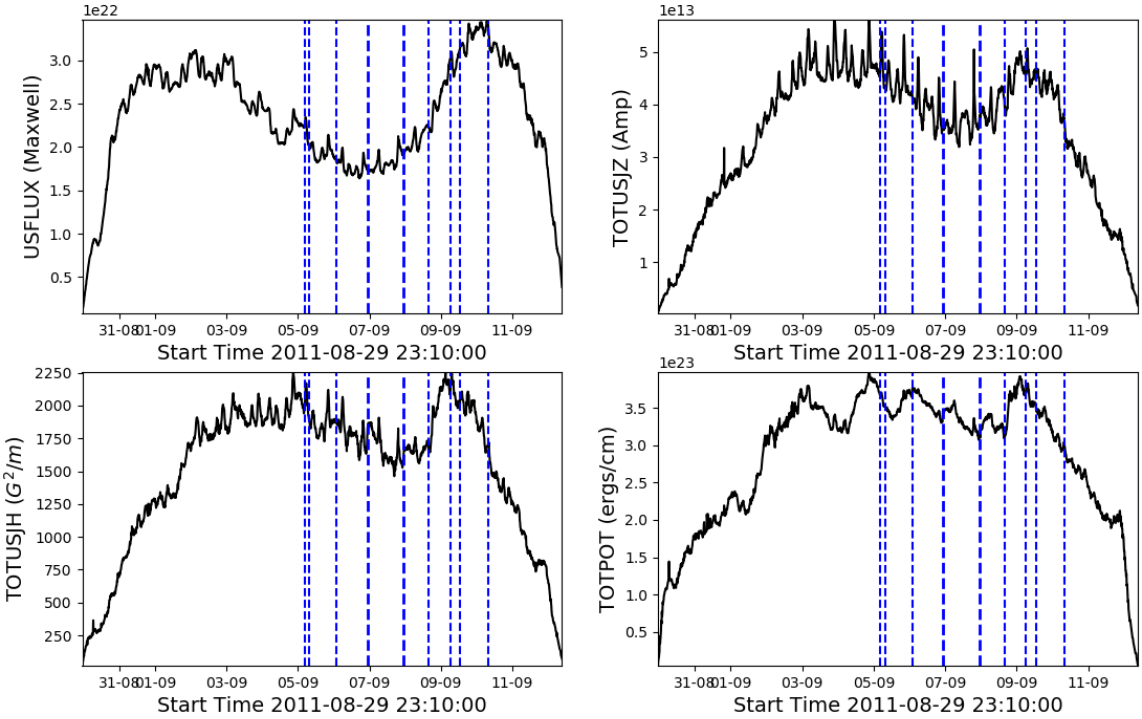

Fig. 2. Time evolution of the total unsigned magnetic flux (USFLUX), the total unsigned vertical current (TOTUSJZ), the total unsigned current helicity (TOTUSJH), and the total photospheric magnetic free-energy density (TOTPOT) for active region NOAA 11283 (SHARP 833). The thick dashed vertical blue line corresponds to an X-class solar flare, and the thin dashed vertical blue line corresponds to an M-class solar flare.

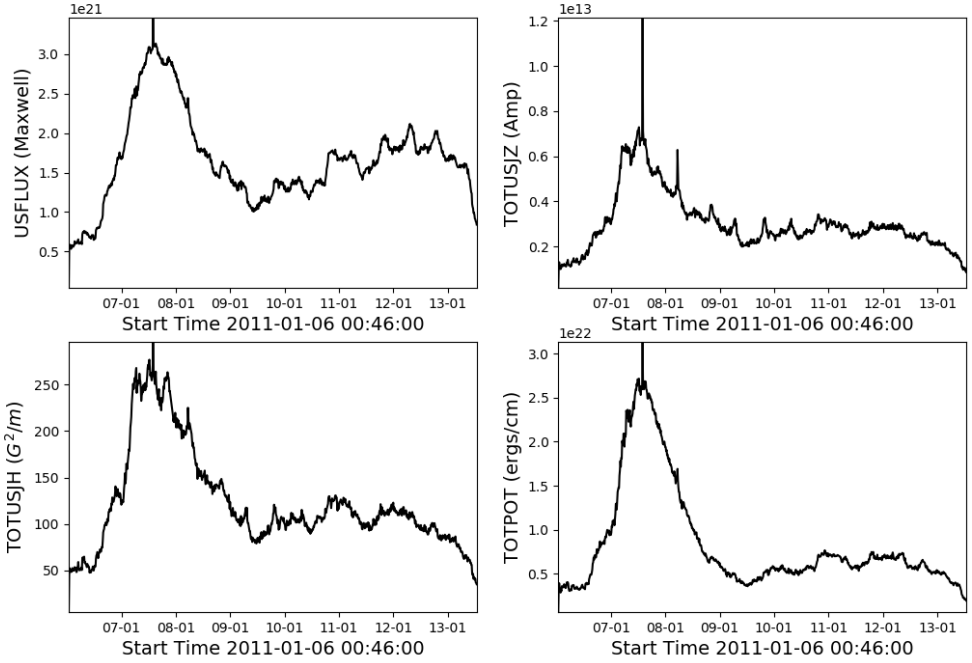

Fig. 3. Time evolution of the total unsigned magnetic flux (USFLUX), the total unsigned vertical current (TOTUSJZ), the total unsigned current helicity (TOTUSJH), and the total photospheric magnetic free-energy density (TOTPOT) for active region NOAA 11143 (SHARP 335). This active region did not produce any solar flare during its transit across the visible solar disk.

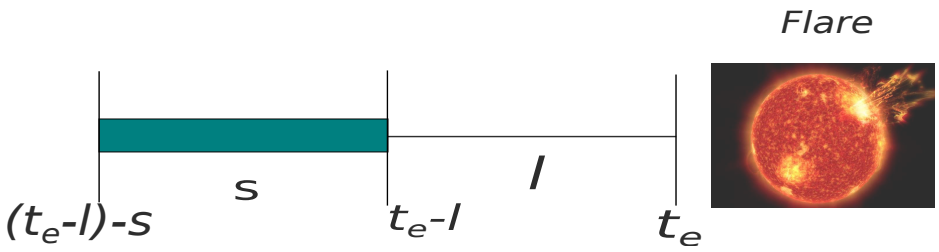

Fig. 4. Data selection criteria for eruptive active regions in terms of flare-generating time (t_e), loopback time (l), and span time (s). We use the time evolution of the active region parameters during span time (s) to train the machine-learning algorithm. The loopback time is also the forecasting window.

Most previous flare-prediction studies considered only magnetic parameters that were 24 h before the flare time. They did not consider any time evolution of the magnetic parameters for their prediction purposes. However, most of the theoretical models suggest a change in magnetic parameters before the solar flare. The standard flare model indicates that flux cancellation near the polarity inversion line is an important determinant for solar flare (van Ballegooijen & Martens 1989). Thus it is necessary to include the basic essence of the time evolution of magnetic parameters in the training purpose.

We considered the time evolution of the magnetic parameters during a time window, named span, for the training purpose. The span time window is always before the loopback time. The loopback is the time window before the occurrence of the solar flare. Figure 4 represents this selection graphically. We assume that we wish to predict a solar flare 24 h before its occurrence, and for this purpose, we considered the 12 h of the time evolution of the magnetic parameters, which is 24 h before the flare occurrence. In this situation, loopback is 24 h and span is 12 h. As we consider 17 magnetic parameters for our study, we obtain 17 time series of the magnetic parameter evolution for the training purpose. We represent each time series of the magnetic parameter evolution by seven statistical parameters associated with the time series: the mean, median, skewness, kurtosis, standard deviation, and the first and third quartile. For a time series named $T = [x_1, x_2, x_3, \ldots, x_n]$, the statistical summary parameters are defined in the following way:

$$\text{Mean } (\mu) = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

$$\text{Std. Deviation } (\sigma) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu(T))^2}{N}} \tag{2}$$

$$\text{Skewness} = \frac{\sum_{i=1}^{n}(x_i - \mu(T))^3}{N\sigma(T)^3} \tag{3}$$

$$\text{Kurtosis} = \frac{\sum_{i=1}^{n}(x_i - \mu(T))^4}{N\sigma(T)^4}. \tag{4}$$

The median is the middle number of the ascending time series T, the first quartile is the middle value between the median and the smallest number of the time series T, and the third quartile is the middle value between the median and the largest number of the time series T. The standard deviation (σ) represents the dispersion around the mean. Skewness and kurtosis are a statistical measure to describe the distribution. While skewness is the measure of the symmetry of the dataset, kurtosis tells us how the tails of the distribution differ from the tails of the normal distribution. We considered 17 magnetic parameters for our study and therefore have a time series for each of the 17 magnetic parameters. As we represent each of the time series by seven statistical parameters, we have 119 entries in the resulting matrix.

4.2. Different supervised machine-learning techniques

There are different supervised machine learning classifiers in the literature. These supervised machine learning algorithms are used for training. Some of the well-known supervised machine-learning classifiers are logistic regression, a decision tree, K nearest neighbors, naive Bayes, a support vector machine, a multilayer perceptron, or a random forest. In the machine-learning literature, every problem is unique. We do not know what algorithms to use or whether the problem can be modeled effectively. A baseline model is the simplest possible prediction model. The result of the baseline model will tell us whether a more complex algorithm adds any value to the result. There is no need for a complex machine-learning algorithm for a particular problem if a simple baseline algorithm can do the same. We here used the logistic regression classifier as a baseline model and compared the baseline result with the results obtained from some other complex machine-learning algorithms, the support vector machine, a random forest, naive Bayes, and a multilayer perceptron.

Baseline model: Logistic regression. Logistic regression (LR) is one of the simplest and most commonly used machine-learning algorithms for the binary classification problem. It is easy to implement, easy to interpret, efficient, and does not require high computation power; thus it can be used as a baseline model for the binary classification problem. This model estimates the probability of an event occurrence by fitting data to a logistic function. The equation used for logistic regression is $\log(\frac{p}{1-p}) = \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_n x_n$, where p is the probability of the event occurrence. $x_1, x_2, \ldots x_n$ is the number of independent variables. $\frac{p}{1-p}$ is known as the odd ratio. If the odd ratio is positive, then the probability of an event occurrence is more than 50%. One of the main drawbacks of the algorithm is the assumption of linearity between the dependent and independent variable. This algorithm separates the classes by constructing a linear decision boundary between them. It is a linear classifier. It does not perform well if the classes are not linearly separable. We used the *lbfgs* solver for our logistic regression classifier. However, a kernel trick can be used to change any linear decision boundary algorithm into a nonlinear decision boundary algorithm.

Support vector machine. The support vector machine (SVM) is a classification algorithm that separates the data of two classes by finding a line (in 1D) or a hyperplane (in higher dimensions) between two classes. In the SVM algorithm, the points near the line or the hyperplane are called support vector, and the distance between the support vector and the line or the hyperplane is called margin. This algorithm tries to find the hyperplane or line by maximizing the margin. SVM is highly suitable for linear classification problems. However, SVM can also solve nonlinear classification problems by moving lower-dimensional space to higher-dimensional space such that we can find the separating hyperplane in the higher dimension. These transformations are known as the kernel trick.

We assumed N training points where each input x_i has D attributes and belongs to any of the two classes $y_i = +1$ or -1. In most of the cases, different classes cannot be solved fully linearly. In this situation, the soft-margin SVM algorithm is commonly used, where the concept of a slack variable and the idea of a trade-off between the minimization of misclassification rate and the maximization of margin is introduced. The hyperplane can be described by the equation $w_i x_i + b - 1 = 0$, where w is the normal to the hyperplane and $b/\|w\|$ is the normal distance from the origin to the hyperplane. In the SVM scenario, $1/\|w\|$ is the margin. In the soft-margin SVM algorithm, we have to select the variable b and w in a way that we can describe our training data by

$$x_i.w + b \geq +1 - \psi_i \text{ for } y_i = +1$$
$$x_i.w + b \leq -1 + \psi_i \text{ for } y_i = -1,$$

where, $\psi_i \geq 0$ is the slack variable. We can combine these two equation into a single equation:

$$y_i(x_i.w + b) - 1 + \psi_i \geq 0.$$

In the soft-margin SVM algorithm, we have to maximize the margin and also reduce the misclassification rate. This can be made by minimizing an objective function subject to the

previous condition. In a more general way, this problem can be defined as

$$\min\left(\frac{1}{2}w^T w + C\sum_{i=1}^{L}\psi_i\right), \tag{5}$$

such that $y_i(w^T\Phi(x_i)+b)-1+\psi_i \geq 0$. Parameter C controls the trade-off between the size of the margin and the slack variable. This parameter is also known as the regularization parameter. Φ is the function that maps the input data into higher dimensional space, also known as the kernel function. It is sometimes difficult to find an appropriate kernel for a particular problem. We do not know whether our problem is linearly separable. We used the kernel trick for our problem. It has previously been shown that a radial basis function kernel projects vectors into infinite-dimensional space. Motivated by this fact, we used a radial basis function kernel for our study. Our choice of the rbf kernel ensures that our support vector machine algorithm will separate the classes by constructing a nonlinear decision boundary.

Multilayer perceptron. The multilayer perceptron (MLP) uses the concept of a neural network to predict the output based on some input features. A perceptron is a linear classifier that separates the input into two classes by a straight line. The output of a perceptron depends on the input, that is, on the feature vector (x), which is multiplied by a weight w and added to a bias b (simply, output = $w.x + b$). The final prediction will be made after passing the output of the perceptron through a nonlinear activation function.

An MLP is a deep neural network. It consists of an input layer where the feature vector is fed, of an output layer for making the prediction about the input, and an arbitrary number of hidden layers in between the input and output layer. Each neuron in the multilayer perceptron is connected with all other neurons of the previous layer. A neuron is a processing unit where inputs are summed using weights and the result is passed through an activation function. In summary, the output of each basic processing unit (neuron) is

$$y = \Phi\left(\sum w_i x_i + b\right) = \Phi(w^T x + b), \tag{6}$$

where x denotes the vector of inputs, w is the vector of weights, b represents the bias, and Φ is the activation function. We used the RELU activation function for each neuron in the hidden layers, and in the output layer, we used a sigmoid activation function.

Training the MLP algorithm involves the adjustment of bias and weights to minimize the error in the output. This is achieved by using forward and backward propagation. MLP is a feed-forward network that involves constant backward and forward propagation until we achieve the desired result.

Forward propagation. In this method, we move the signal from the input layer to the output layer through hidden layers. We measure the output or decision of the output layer with respect to the ground-truth label. This measurement is also known as error.

Backward propagation. In this process, we back-propagate the partial derivative of the error with respect to weights and bias from the output layer. A stochastic gradient descent algorithm is used to adjust the weights and bias in this process.

The multilayer perceptron model has some advantage. A very complex model can be trained by the MLP model, and no feature engineering is required before training. However, it is difficult to explain the MLP model simply, and parameterization is also complex. This model also needs more training data.

Random forest. This algorithm is an ensemble algorithm that is used for both regression and classification. This algorithm uses the aggregate of multiple decision trees to make the prediction. This algorithm first selects some random bootstrap samples from a given dataset. Next, it generates decision trees for every random sample and obtains the prediction from each decision tree. Finally, it gives the final prediction based on the voting of each predicted result obtained from multiple decision trees. This model has one major advantage. It overcomes the problem of overfitting by averaging the results. Complexity and more computing resource requirement is the main disadvantage of this algorithm.

K-nearest neighbor and naive Bayes classifiers. The K-nearest neighbor (KNN) algorithm classifies the samples based on a similarity measure (e.g., distance function). It classifies the case by the majority vote of nearest neighbors. This algorithm is easy to implement and easy to interpret. However, the KNN algorithm has dimensionality problems. It does not perform well with a large number of features.

Naive Bayes is a probability-based algorithm that separates the classes based on the Bayes theorem. It calculates the probabilities for a particular class for given features. One main assumption made in this algorithm is that features are independent of each other. This is why the algorithm is called naive. Finally, this algorithm considers the class with the highest probability as the most likely class.

4.2.1. Performance measure and class imbalance problem

We obtain a confusion matrix as a result of our classification algorithm that consists of four entries: TP, TN, FP, and FN. Here TP (true positive) are the cases where positively labeled samples are correctly predicted as positive, TN (true negative) are the cases where negatively labeled samples are correctly predicted as negative, FP (false positive) are the cases where negatively labeled samples are predicted as positive, and FN (false negative) are the cases where positively labeled samples are incorrectly predicted as negative. Accuracy in the classification problem is defined as the number of correct predictions made by the model over the total number of predictions made:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}. \tag{7}$$

Accuracy is a good performance measure when the data set is balanced. Some other performance measures are precision ($TP/(TP+FP)$), recall ($TP/(TP+FN)$), and the F-score (harmonic mean of precision and recall).

As solar active regions do not have flares or an eruption most of the time, there are more nonflaring active regions than flaring active regions. There is a huge imbalance between the number of flaring and nonflaring active regions. This problem is known as the class imbalance problem. In this case, accuracy will be very high if the model predicts almost all active regions as nonflaring (as the number of nonflaring regions is very high). However, we aim to predict the flaring active regions, which are rare. Thus, accuracy is not a good performance measure for the class imbalance problem. Later, some other performance measures, such as the Heidke skill score (HSS_1 and HSS_2) and the Gilbert score (GS) are used (Barnes & Leka 2008; Mason & Hoeksema 2010;

Ahmadzadeh et al. 2019):

$$\mathrm{HSS}_1 = \frac{TP + TN - N}{P} \tag{8}$$

$$\mathrm{HSS}_2 = \frac{2 \times [(TP \times TN) - (FN \times FP)]}{P \times (FN + TN) + (TP + FP) \times N} \tag{9}$$

$$\mathrm{GS} = \frac{TP \times (P + N) - (TP + FP) \times (TP + FN)}{(TP + FP + FN) \times (P + N) - (TP + FP) \times (TP + FN)}, \tag{10}$$

where P and N are the total number of actual positive and negative samples. HSS_1 measures the improvement of the prediction over all negative predictions, while HSS_2 measures the same, but over a random forecast. The Gilbert score (GS) measures the number of TPs obtained by chance.

However, all these measures HSS_1, HSS_2 and GS have some dependency on the ratio of class imbalance. To alleviate this problem, Bloomfield et al. (2012) introduced a new performance measure, the true skill statistic (TSS), which is independent of the class imbalance ratio. True skill statistic is defined as

$$\mathrm{TSS} = \frac{TP}{TP + FN} - \frac{FP}{TN + FP}. \tag{11}$$

The TSS value varies from −1 to +1, where perfect correct prediction scores +1, always incorrect prediction scores −1, and random prediction scores zero. Flare prediction is a highly imbalanced problem. The TSS is the most meaningful measure in the case of the flare prediction scenario as it does not depend on the ratio of the class imbalance. The code for this study can be found in our github repository[2].

4.2.2. Results

We use active region vector magnetogram data from June 2010 to December 2017 for this study. After generating the dataset, we performed some preprocessing in our data set because some values were missing. We replaced these with the mean of the corresponding features. It was also necessary to normalize the dataset before training because it will transform the ranges of all feature values into a uniform range. We used the zero-one data transformation technique for normalization. We randomly divided our dataset into a training set (70%) and a testing data set (30%). We maintained the same class ratio (N/P) in the training and testing data set following the prescription of Bobra & Couvidat (2015) and Ahmed et al. (2013). We did not include C-class flares in our positive data set. Our data set is highly imbalanced because there are far fewer flaring than nonflaring regions. Oversampling and undersampling are the two well-known strategies to make an imbalanced data set balanced. While oversampling involves the strategy of adding more positive examples in the data set, undersampling involves the strategy of removing most of the negative examples from the data set. However, both methods have some limitations. In oversampling, the addition of many replicas of positive examples may cause the model to memorize the pattern; this causes the model to be prone to overfitting. On the other hand, we removed many negative examples in the undersampling strategy, therefore the computer did not learn from the entire data set.

We used the concept of weighted techniques following Ahmadzadeh et al. (2019) to solve the issue of class imbalance. A detailed investigation of the class imbalance issue is beyond the scope of our current study. In the weighted techniques, we provided more weights to the classes that we aim to predict. The weighted technique is also free from the limitations of undersampling and oversampling techniques. We aimed to predict flaring active regions, of which there are fewer. We provided more weight to the flaring active regions than to nonflaring regions such that our classifier did not exclusively focus on the nonflaring class, which is the majority one. In our SVM classifier, we set the cost parameter of the flaring class higher than that of the nonflaring class to solve the issue of class imbalance. We also used a similar technique for our baseline classifier to address the issue of class imbalance. In the case of our other classifier, MLP, we used the weight-balancing technique to solve the issue of class imbalance. In the weight-balancing technique, we altered the weights of each training data during the computation of loss function. Generally, both positive and negative classes carry the same weight 1.0 in the computation of loss function. As our primary aim here is to predict minority classes (flaring one), we provided more weight to the flaring classes in the calculation of the loss function than to the non-flaring one.

Table 2 lists the performance metrics found after running the baseline-weighted logistic regression classifier on the five different data sets. We provide the means and standard deviations in the table by repeating training and testing phases several times. We used a 24 h forecasting window for the data sets loop24span12, loop24span24, and loop24span0. We used a 12 and 24 h time evolution of the magnetic parameters for the data sets loop24span12 and loop24span24 as training purposes (spans of 12 h and 24 h). We used no time evolution of the magnetic parameters for the data set loop24span0 for training purposes. The loop24span0 data set is almost similar to that of Bobra & Couvidat (2015), but it covers a longer time (June 2010 to December 2017). We used a 12 h time evolution of the magnetic parameters for the data sets loop12span12 and loop48span12 for training purposes, but we used different forecasting windows. The forecasting window was 12 h for the data set loop12span12, and the forecasting window for the data set loop48span12 was 48 h. We found a very good TSS value for all five cases (see Table 2). We obtained the maximum TSS value (0.92) for the data set loop12span12. Interestingly, we also find a very good TSS value for the data set loop24span0, where we did not use any time evolution for training. In summary, the time evolution does not have much effect on the performance of a classifier. We also find a good TSS value for the data set loop48span12. Our baseline model result indicates that it is possible to distinguish a flaring active region from a nonflaring one.

Table 3 shows the different performance metrics we found after running some other classifiers such as the SVM, the MLP, the RF, KNN, and naive Bayes on the loop24span12 data set. We also list the results of our baseline classifier for comparison. We tuned the hyperparameters of the SVM algorithm by using a grid-search algorithm provided by the Scikit-Learn software package (Pedregosa et al. 2012). This algorithm finds the best combination of hyperparameters after performing an exhaustive search over a predefined set of hyperparameters. Our SVM hyperparameter, regularization parameter (C), varies between 0.001 and 10, and the RBF parameter γ varies between 0.001 and 1. We used *binary cross-entropy* as a loss function and *rmsprop* as an optimizer for our multilayer perceptron algorithm. We also used the *Keras l2* regularizer as a kernel regularizer for our MLP algorithm. The hyperparameters for our other classifiers are the following: RF (n estimators = 100, max depth = none, criterion = gini, class weight = balanced), naive Bayes (priors = none), and KNN (number of neighbors $k = 1$,

[2] http://gitub.com/soumitrahazra/Flaring_Region_Prediction

Table 2. Flare prediction capabilities obtained from our baseline model LR for five different data sets.

Results from our baseline model: Logistic regression						
		loop24span12	loop24span24	loop24span0	loop12span12	loop48span12
Considering all magnetic parameters	Accuracy	0.95 ± 0.023	0.94 ± 0.007	0.94 ± 0.008	0.96 ± 0.006	0.92 ± 0.010
	Precision (Positive)	0.84 ± 0.023	0.84 ± 0.019	0.79 ± 0.026	0.87 ± 0.021	0.77 ± 0.025
	Precision (Negative)	0.98 ± 0.005	0.98 ± 0.005	0.98 ± 0.005	0.97 ± 0.003	0.97 ± 0.006
	Recall (Positive)	0.93 ± 0.018	0.92 ± 0.016	0.93 ± 0.019	0.96 ± 0.014	0.89 ± 0.021
	Recall (Negative)	0.95 ± 0.008	0.96 ± 0.006	0.94 ± 0.010	0.97 ± 0.004	0.92 ± 0.011
	F_1-score (Positive)	0.88 ± 0.014	0.88 ± 0.013	0.85 ± 0.017	0.91 ± 0.014	0.83 ± 0.018
	F_1-score (Negative)	0.97 ± 0.004	0.97 ± 0.004	0.96 ± 0.005	0.97 ± 0.004	0.94 ± 0.007
	HSS_1	0.75 ± 0.031	0.74 ± 0.033	0.69 ± 0.041	0.81 ± 0.025	0.64 ± 0.039
	HSS_2	0.85 ± 0.018	0.85 ± 0.019	0.82 ± 0.022	0.89 ± 0.015	0.77 ± 0.023
	GS	0.74 ± 0.026	0.73 ± 0.028	0.68 ± 0.030	0.80 ± 0.024	0.63 ± 0.030
	TSS	0.87 ± 0.016	0.87 ± 0.016	0.86 ± 0.018	0.92 ± 0.014	0.81 ± 0.023
Considering only the top five magnetic parameters	Accuracy	0.94 ± 0.006	0.93 ± 0.007	0.94 ± 0.006	0.96 ± 0.005	0.91 ± 0.009
	Precision (Positive)	0.83 ± 0.023	0.81 ± 0.022	0.83 ± 0.022	0.85 ± 0.020	0.78 ± 0.025
	Precision (Negative)	0.97 ± 0.006	0.97 ± 0.006	0.96 ± 0.005	0.98 ± 0.005	0.95 ± 0.008
	Recall (Positive)	0.88 ± 0.023	0.87 ± 0.024	0.86 ± 0.023	0.91 ± 0.021	0.84 ± 0.027
	Recall (Negative)	0.95 ± 0.008	0.95 ± 0.007	0.96 ± 0.004	0.96 ± 0.006	0.93 ± 0.011
	F_1-score (Positive)	0.85 ± 0.015	0.84 ± 0.017	0.84 ± 0.014	0.88 ± 0.014	0.81 ± 0.018
	F_1-score (Negative)	0.96 ± 0.004	0.96 ± 0.005	0.96 ± 0.004	0.97 ± 0.003	0.94 ± 0.006
	HSS_1	0.69 ± 0.035	0.68 ± 0.036	0.68 ± 0.034	0.75 ± 0.029	0.61 ± 0.041
	HSS_2	0.81 ± 0.021	0.81 ± 0.022	0.80 ± 0.019	0.85 ± 0.017	0.75 ± 0.024
	GS	0.68 ± 0.029	0.67 ± 0.031	0.67 ± 0.028	0.74 ± 0.025	0.60 ± 0.031
	TSS	0.83 ± 0.025	0.82 ± 0.023	0.82 ± 0.021	0.87 ± 0.017	0.76 ± 0.027

Notes. The first three data sets, loop24span12, loop24span24, and loop24span0, correspond to the data sets with the same forecasting window of 24 h, but a different span time of 12 h, 48 h, and zero hours, respectively. The last two data sets, loop12span12 and loop48span12, correspond to the data sets with the same span time of 12 h, but have a different forecasting window of 12 h and 48 h, respectively.

Table 3. Flare prediction capabilities obtained from different classifiers for the loop24span12 data set.

Performance by other classifiers on our loop24span12 data set						
Our classifiers:	LR	SVM	MLP	KNN	Random forest	Naive Bayes
Accuracy	0.95 ± 0.023	0.96 ± 0.019	0.96 ± 0.017	0.95 ± 0.006	0.96 ± 0.006	0.94 ± 0.008
Precision (Positive)	0.84 ± 0.023	0.90 ± 0.022	0.86 ± 0.080	0.90 ± 0.022	0.93 ± 0.019	0.82 ± 0.033
Precision (Negative)	0.98 ± 0.005	0.98 ± 0.005	0.98 ± 0.009	0.97 ± 0.006	0.97 ± 0.005	0.97 ± 0.006
Recall (Positive)	0.93 ± 0.018	0.92 ± 0.019	0.94 ± 0.036	0.85 ± 0.023	0.90 ± 0.020	0.88 ± 0.024
Recall (Negative)	0.95 ± 0.008	0.97 ± 0.006	0.95 ± 0.031	0.97 ± 0.005	0.98 ± 0.005	0.95 ± 0.012
F_1-score (Positive)	0.88 ± 0.014	0.91 ± 0.014	0.89 ± 0.032	0.88 ± 0.017	0.91 ± 0.013	0.85 ± 0.017
F_1-score (Negative)	0.97 ± 0.004	0.96 ± 0.005	0.97 ± 0.016	0.97 ± 0.004	0.97 ± 0.008	0.96 ± 0.005
HSS_1	0.75 ± 0.031	0.83 ± 0.028	0.77 ± 0.016	0.76 ± 0.034	0.83 ± 0.023	0.69 ± 0.040
HSS_2	0.85 ± 0.018	0.89 ± 0.017	0.86 ± 0.054	0.85 ± 0.020	0.89 ± 0.015	0.81 ± 0.022
GS	0.74 ± 0.026	0.81 ± 0.028	0.76 ± 0.071	0.74 ± 0.032	0.80 ± 0.024	0.68 ± 0.030
TSS	0.87 ± 0.016	0.90 ± 0.018	0.89 ± 0.029	0.83 ± 0.024	0.88 ± 0.019	0.83 ± 0.020

distance = Eucledian). A comparison between the TSS values obtained from different classifiers confirms the robustness of our model in distinguishing flaring and nonflaring active regions. We also note that our baseline classifier LR (which is easy to interpret) works very well in distinguishing flaring and nonflaring regions. Although a detailed investigation about the class imbalance issue is beyond the scope of the current study, here we have provided a comparison study between the results obtained from the classifiers with and without class remedy. Table 4 shows the confusion matrices obtained from the classifiers with and without class remedy. We note that FN decrease when we take the remedy of class imbalance in our model into account. The reduction of FN is more important than the reduction of FP in the case of the forecasting study. In other words, recall is also an important metric. In summary, we find that the performance of the classifier improves when we perform the remedy of class imbalance.

The question now is which magnetic parameter is most critical for distinguishing flaring and nonflaring active regions. We followed the suggestions of Hamdi et al. (2017) to determine the best active region parameter. We used the time evolution of the magnetic parameters for training machine-learning algorithms. The magnetic parameter whose corresponding time evolution gives the highest TSS after the classification by the LR (baseline) and SVM algorithm is considered to be the best of all active region magnetic parameters in terms of distinguishing capability. We used the data set loop24span12 for this experiment. For this

Table 4. Confusion matrices obtained from our three different models for the loop24span12 data set.

Our models:	TP	FP	TN	FN
Logistic regression (without class imbalance remedy)	163	11	750	30
Logistic regression (with class imbalance remedy)	177	27	734	16
Multilayer perceptron (with class imbalance remedy)	190	54	707	3

purpose, we trained the SVM and LR algorithms by using the time evolution of a single magnetic parameter at a time and measured the TSS value for each case. The bar plot of Fig. 5 shows that the total unsigned magnetic helicity and the total unsigned magnetic flux achieve maximum TSS. These results indicate that the time evolution of the total unsigned magnetic helicity and the total unsigned magnetic flux is the best indicator in terms of distinguishing capability. Bobra & Couvidat (2015) also found that the total unsigned magnetic helicity parameter is the best active region parameter based on the Fisher criterion. However, we find that the time evolution of the total unsigned magnetic flux has an equally good distinguishing capability. We also note that the time evolution of some other parameters, the total photospheric magnetic free energy density, the total unsigned vertical current, and AREA_ACR, have a good distinguishing capability as well (see Fig. 5). The time evolution of the total unsigned magnetic flux is an indicator of flux cancellation and emergence on the solar surface. On the other hand, the time evolution of the magnetic helicity indicates the magnetic complexity of the active region. Our finding in terms of critical active region parameter is consistent with the earlier theoretical and observational findings.

Previous studies suggested that there will be no overfitting if we use 12 to 18 magnetic parameters (Bobra & Ilonidis 2016; Inceoglu et al. 2018). Because of this, we did not apply any feature selection criterion before applying a machine-learning algorithm. Now, we selected only the top 5 magnetic parameters for our study. These are the total unsigned magnetic flux (USFLUX), the total unsigned current helicity (TOTUSJH), the total unsigned vertical current (TOTUSJZ), the total photospheric magnetic free energy density (TOTPOT), and the area of strong-field pixels in the active region (AREA_ACR). We trained our baseline logistic regression algorithm using the time evolution of these five magnetic parameters. The bottom part of Table 2 lists the performance metrics found after running the classifiers over the data sets. It considers only the best five magnetic parameters. We find that our baseline classifier performs quite well in terms of distinguishing capability even if we consider only the top five magnetic parameters. We also note that TSS values obtained using only the top five magnetic parameters are very close to the values obtained by considering all magnetic parameters.

To investigate how the classifier performance changes with forecast window, we created a data sets with different forecasting windows (loop) but the same span. Table 2 shows the classification metrics obtained after running the classifier over three data sets, loop12span12, loop24span12, and loop48span12. We used a 12 h time evolution of the magnetic parameters for training in all cases. Table 2 shows that there is a decreasing trend of the TSS value with our selected forecasting windows. However, we obtain very good TSS values for all three data sets. Our selected forecasting windows affect the performance of the classifier very little.

We considered data sets with a fixed forecasting window of 24 h but different span windows to determine how the time evolution of magnetic parameters for different time windows (span) changes the performance of the classifier. We did not find any increasing or decreasing trend of the TSS value with the span size. The span size may not have much effect on the performance of the classifier.

5. Summary

We have performed a comparison study between eruptive and noneruptive active regions in terms of the time evolution of the photospheric magnetic parameters. We first performed this study manually. We selected two eruptive active regions and one noneruptive active region to determine the difference between the time evolution of the magnetic complexity in eruptive and noneruptive active regions. We find a significant difference between the eruptive and noneruptive active region in terms of both strength and time evolution of the photospheric magnetic parameters. All of our selected magnetic parameters, that is, the total unsigned magnetic flux, the total unsigned helicity, the total photospheric magnetic free-energy density, and the total unsigned vertical current, have a much higher value in case of eruptive active regions than in noneruptive active region. We find the signature of flux emergence and cancellation in the time evolution of the total unsigned magnetic flux. The time evolution of the total unsigned current helicity shows a strong indication of both helicity accumulation and destruction at the onset of a solar flare.

As it is not possible to analyze all flaring events manually because of the large amount of solar data, we used machine-learning algorithms to distinguish eruptive active regions from the noneruptive regions. We used the time evolution of the photospheric magnetic parameters to train our baseline machine-learning algorithm logistic regression. Motivated by Bloomfield et al. (2012) and Bobra & Couvidat (2015), we selected the TSS as a performance measure of our forecasting algorithm because the TSS does not depend on class imbalance. Solar flare prediction is a highly imbalanced problem as there is a fewer flaring active region compared to non-flaring ones. We obtain high TSS values for our baseline algorithm. A higher TSS value also implies a lower false-negative rate, which is very important for the forecasting study. We then compared our baseline result with the results obtained from our other classifiers. We find that our model results are robust in terms of distinguishing capability.

We note that our TSS value is higher than that in previous studies. This may be due to our data set (because the data set is different from previous studies) or it may be due to our choice of training the computer by the evolution of the photospheric magnetic parameters. Martens & Angryk (2018) described the importance of the "benchmark data set" for this type of prediction studies. Preparation of the benchmark data set involves many processes such as gathering a large amount of data, that then need to be cleaned and balanced. We also performed one study where we did not use the time evolution of the photospheric magnetic parameters to train the machine-learning algorithm. We found a very good TSS value in that case. Therefore we are not sure whether the time evolution of the magnetic parameters is really important for a flare-forecasting study. However, we described the time evolution by eight statistical parameters, which may not be a very good way to describe the time evolution (but it is the simplest way). We did not consider any statistical quantities to represent seasonal or periodic features in

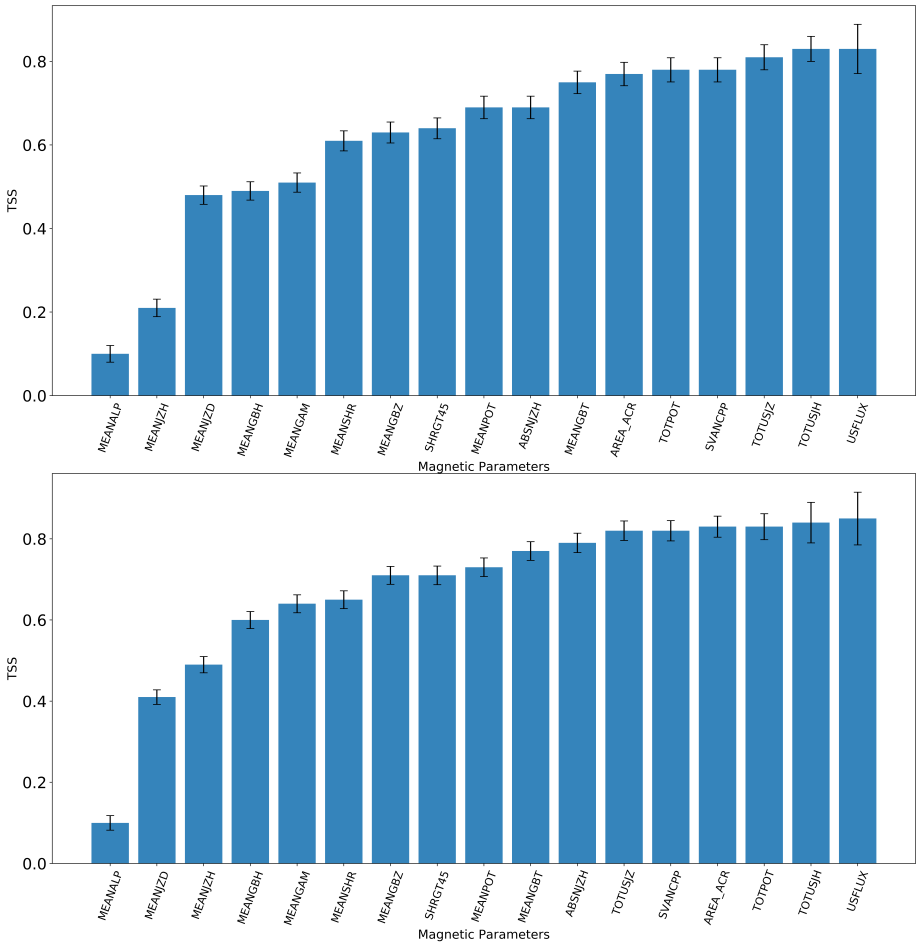

Fig. 5. *Top bar plot*: distribution of TSS values after running the logistic regression (baseline) classifier on the summarized time series of individual active region parameters. *Bottom bar plot*: same, but for the support vector machine algorithm. This plots clearly show that the top few parameters, specifically, the total unsigned magnetic flux and the total unsigned magnetic helicity, have the best distinguishing capability.

this study. Features generated from fast Fourier transform (FFT) of the time series will be helpful to capture the information of periodic features of the time series. Different characteristics of the time series carry some important pieces of information. There is also no established work at this moment that clearly determines the effective statistical features for improved flare forecasting. We leave this investigation for our future studies.

Next, we determined the common critical magnetic parameter that clearly distinguishes a eruptive active regions from a noneruptive region. We find that the time evolution of the total unsigned magnetic helicity and the total unsigned magnetic flux have a very high distinguishing capability. We note that the time evolution of the unsigned magnetic flux is an indicator of flux emergence and cancellation on the solar surface, while the time evolution of the unsigned magnetic helicity represents the helicity accumulation and destrucitonn. Previous theoretical models described the importance of these two mechanisms in detail at the onset phase of the solar flare. Our manual analysis with three active regions also supports this view. Now, we compare this result with some earlier studies. Bobra & Couvidat (2015) found that the total unsigned magnetic helicity has the best distinguishing capability compared to others. However, they did not find any significant distinguishing capability of the total unsigned magnetic flux. They did not use the time evolution of the magnetic parameters to train a machine-learning algorithm. Ma et al. (2017) used the univariate time-series clustering and multivariate time-series decision tree for the purpose of flare prediction and found that both the total unsigned magnetic flux and the total unsigned magnetic helicity have very high distinguishing capability compared to other parameters.

We also find a very high TSS value when we considered only the evolution of the top five magnetic parameters for the training of machine-learning algorithms. In summary, although we are not able to find a single critical magnetic parameter, we find that a combination of the top few magnetic parameters will give us almost similar distinguishing capability. This result is consistent with earlier studies. Earlier studies also found that the top few magnetic parameters can produce a forecasting capability comparable to their entire data set (Leka & Barnes 2003, 2007; Ahmed et al. 2013; Bobra & Couvidat 2015; Hamdi et al. 2017). We note that "total" parameters are more valuable in a flare study compared to the mean parameters. Welsch et al. (2009) also found that extensive magnetic parameters (whose value increases with size) have a stronger correlation with the flare productivity than the intensive parameters (value does not increase with size). This result indicates that larger and complex active regions are more flare prone than a smaller region.

Acknowledgements. The Center of Excellence in Space Sciences India (CESSI) is supported by the Ministry of Human Resource Development, Government of India. We thank the anonymous referee for valuable comments and suggestions. We thank Georgia State University data mining group for the discussion and suggestion. We also thank Shaun Bloomfield for reading the manuscript and providing valuable suggestions.

References

Abramenko, V. I., Wang, T., & Yurchishin, V. B. 1996, Sol. Phys., 168, 75
Aggarwal, A., Schanche, N., Reeves, K. K., Kempton, D., & Angryk, R. 2018, ApJS, 236, 15

Ahmadzadeh, A., Hostetter, M., Aydin, B., et al. 2019, ArXiv e-prints [arXiv:1911.09061]
Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, Sol. Phys., 283, 157
Amari, T., Aly, J. J., Mikic, Z., & Linker, J. 2010, ApJ, 717, L26
Ambastha, A., Hagyard, M. J., & West, E. A. 1993, Sol. Phys., 148, 277
Barnes, G., & Leka, K. D. 2008, ApJ, 688, L107
Berger, M. A., & Field, G. B. 1984, J. Fluid Mech., 147, 133
Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJ, 747, L41
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., & Ilonidis, S. 2016, ApJ, 821, 127
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, Sol. Phys., 289, 3549
Burtseva, O., & Petrie, G. 2013, Sol. Phys., 283, 429
Crouch, A. D., Barnes, G., & Leka, K. D. 2009, Sol. Phys., 260, 271
Dhuri, D. B., Hanasoge, S. M., & Cheung, M. C. M. 2019, Proc. Natl. Acad. Sci., 116, 11141
Filali Boubrahimi, S., & Angryk, R. 2018, 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 162
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, Sol. Phys., 293, 28
Gaizauskas, V., Zirker, J. B., Sweetland, C., & Kovacs, A. 1997, ApJ, 479, 448
Hamdi, S. M., Kempton, D., Ma, R., Boubrahimi, S. F., & Angryk, R. A. 2017, 2017 IEEE International Conference on Big Data (Big Data), 2543
Hazra, S., Nandy, D., & Ravindra, B. 2015, Sol. Phys., 290, 771
Hazra, S., Mahajan, S. S., Douglas, W. K., Jr, & Martens, P. C. H. 2018, ApJ, 865, 108
Holder, Z. A., Canfield, R. C., McMullen, R. A., et al. 2004, ApJ, 611, 1149
Inceoglu, F., Jeppesen, J. H., Kongstad, P., et al. 2018, ApJ, 861, 128
Kusano, K., Suzuki, Y., & Nishikawa, K. 1995, ApJ, 441, 942
Kusano, K., Maeshiro, T., Yokoyama, T., & Sakurai, T. 2002, ApJ, 577, 501
Kusano, K., Yokoyama, T., Maeshiro, T., & Sakurai, T. 2003, Adv. Space Res., 32, 1931
Kusano, K., Maeshiro, T., Yokoyama, T., & Sakurai, T. 2004, ApJ, 610, 537
Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1296
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Leka, K. D., Canfield, R. C., McClymont, A. N., et al. 1993, ApJ, 411, 370
Livi, S. H. B., Martin, S., Wang, H., & Ai, G. 1989, Sol. Phys., 121, 197
Low, B. C. 1977, ApJ, 217, 988
Ma, R., Boubrahimi, S. F., Hamdi, S. M., & Angryk, R. A. 2017, 2017 IEEE International Conference on Big Data (Big Data), 2569
Martens, P. C., & Angryk, R. A. 2018, in Space Weather of the Heliosphere: Processes and Forecasts, eds. C. Foullon, & O. E. Malandraki, IAU Symp., 335, 344
Martens, P. C., & Zwaan, C. 2001, ApJ, 558, 872
Martin, S. F., Livi, S. H. B., & Wang, J. 1985, Max Planck Inst. Astrophys. Rep., 212, 179
Martin, S. F., Bilimoria, R., & Tracadas, P. W. 1994, in NATO Advanced Science Institutes (ASI) Series C, eds. R. J. Rutten, & C. J. Schrijver, 433, 303
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
Metcalf, T. R. 1994, Sol. Phys., 155, 235
Metcalf, T. R., Leka, K. D., & Mickey, D. L. 2005, ApJ, 623, L53
Moon, Y. J., Chae, J., Choe, G. S., et al. 2002, ApJ, 574, 1066
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Park, S.-H., Lee, J., Choe, G. S., et al. 2008, ApJ, 686, 1397
Park, S.-H., Cho, K.-S., Bong, S.-C., et al. 2012, ApJ, 750, 48
Park, S.-H., Kusano, K., Cho, K.-S., et al. 2013, ApJ, 778, 13
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2012, ArXiv e-prints [arXiv:1201.0490]
Pevtsov, A. A., Canfield, R. C., & Metcalf, T. R. 1994, ApJ, 425, L117
Priest, E. R., & Forbes, T. G. 2002, A&ARv, 10, 313
Scherrer, P. H., Schou, J., Bush, R. I., et al. 2012, Sol. Phys., 275, 207
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, Sol. Phys., 275, 229
Schrijver, C. J. 2007, ApJ, 655, L117
Shibata, K., & Magara, T. 2011, Liv. Rev. Sol. Phys., 8, 6
Sinha, S., Srivastava, N., & Nandy, D. 2019, ApJ, 880, 84
Spirock, T. J., Yurchyshyn, V. B., & Wang, H. 2002, ApJ, 572, 1072
Sudol, J. J., & Harvey, J. W. 2005, ApJ, 635, 647
Tian, L., Liu, Y., & Wang, J. 2002, Sol. Phys., 209, 361
Tur, T. J., & Priest, E. R. 1976, Sol. Phys., 48, 89
van Ballegooijen, A. A., & Martens, P. C. H. 1989, ApJ, 343, 971
Vemareddy, P., Ambastha, A., Maurya, R. A., & Chae, J. 2012, ApJ, 761, 86
Wang, H., & Tang, F. 1993, ApJ, 407, L89
Wang, J., Shi, Z., Wang, H., & Lue, Y. 1996, ApJ, 456, 861
Welsch, B. T., Li, Y., Schuck, P. W., & Fisher, G. H. 2009, ApJ, 705, 821
Yu, D., Huang, X., Wang, H., & Cui, Y. 2009, Sol. Phys., 255, 91
Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. 2010, Res. Astron. Astrophys., 10, 785
Zhang, H., & Bao, S. 1999, ApJ, 519, 876
Zirin, H., & Wang, H. 1993, Sol. Phys., 144, 37