RESEARCH

Interpretable ML-Based Forecasting of CMEs Associated with Flares

Hemapriya Raju[1] · Saurabh Das[1]

Abstract

Coronal mass ejections (CMEs) that cause geomagnetic disturbances on the Earth can be found in conjunction with flares, filament eruptions, or independently. Though flares and CMEs are understood as triggered by the common physical process of magnetic reconnection, the degree of association is challenging to predict. From the vector magnetic field data captured by the *Helioseismic and Magnetic Imager* (HMI) onboard the *Solar Dynamics Observatory* (SDO), active regions are identified and tracked in what is known as Space Weather HMI Active Region Patches (SHARPs). Eighteen magnetic field features are derived from the SHARP data and fed as input for the machine-learning models to classify whether a flare will be accompanied by a CME (positive class) or not (negative class). Since the frequency of flare accompanied by CME occurrence is less than flare alone events, to address the class imbalance, we have explored the approaches such as undersampling the majority class, oversampling the minority class, and synthetic minority oversampling technique (SMOTE) on the training data. We compare the performance of eight machine-learning models, among which the Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) model perform best with True Skill Score (TSS) around 0.78 ± 0.09 and 0.8 ± 0.05, respectively. To improve the predictions, we attempt to incorporate the temporal information as an additional input parameter, resulting in LDA achieving an improved TSS of 0.92 ± 0.04. We utilize the wrapper technique and permutation-based model interpretation methods to study the significant SHARP parameters responsible for the predictions made by SVM and LDA models. This study will help develop a real-time prediction of CME events and better understand the underlying physical processes behind the occurrence.

Keywords SHARP parameters · Coronal mass ejections · Class imbalance · Feature importance · Machine learning

1. Introduction

Space weather disturbances are caused due to the interaction of the solar wind and radiation with the Earth's atmosphere (Gosling et al., 1991). Solar sources of such disturbances can be

✉ H. Raju
  hemapriya.rceg@gmail.com

  S. Das
  saurabh.das@iiti.ac.in

[1] Department of Astronomy, Astrophysics and Space Engineering, IIT Indore, Indore, India

sudden and intense such as flares, coronal mass ejections (CMEs) (Priest and Forbes, 2002; Schrijver, 2009; Lin and Forbes, 2000) or stable, and recurrent such as high-speed streams (HSSs) due to open magnetic fields from coronal hole regions (Zirker, 1977; Altschuler, Trotter, and Orrall, 1972; Cranmer, 2009). Understanding the physical mechanisms behind these sudden events and forecasting them in advance is essential for satellite and telecommunication systems. Flares are localized bright eruptions that can be observed as the sudden increase in X-ray flux intensity and release around $10^{28} - 10^{32}$ ergs of energy. CMEs are global phenomena that expel a large volume of plasma to the interplanetary space causing interplanetary shocks, geomagnetic disturbances, and injection of solar energetic particles (SEPs) (Webb and Howard, 2012). During solar maximum, active regions that are characterized by closed, strong magnetic fields are seen very prominently. Reorganization of these magnetic structures due to factors such as flux emergence or shearing causes the magnetic field to become more twisted leading to deviation from its original potential configuration. This, in turn, increases the storage of total magnetic free energy (Shibata and Magara, 2011) in the lower atmosphere corona. After a certain threshold, instability caused by physical mechanisms such as magnetic reconnection results in the release of this stored magnetic energy from the chromosphere and corona in the form of CMEs. Thus identifying those parameters that contribute to an increase in global non-potentiality can be correlated with the occurrence of flares and CMEs.

Flares and CMEs necessarily do not drive each other and can be independent. The underlying physical mechanisms for both events are thought to be the result of the same phenomena caused by magnetic reconnection (Gosling, 1993; Yan, Qu, and Kong, 2011). Harrison (1995) statistically studied the association of CMEs with flares through temporal and spatial analysis of CME onsets observed from white-light images with flare properties such as the flare duration and intensity. They observed that the major flares, such as M class, X class, or subsequent flares, are often associated with CME occurrences. The study carried out by Andrews (2003) for the period of 1996 – 1999, shows that flares with longer durations have a greater chance of being associated with CMEs, varying from 26% for flares lasting less than 2 hours to 100% for flares lasting more than 6 hours. There was no significant observed variation in X-ray peak flux for large flares that were not associated with CMEs. Statistical studies associating CME occurrence rate with flare peak-flux, its fluence, and duration for the period 1996 – 2007 using data from the *Geostationary Operational Environmental Satellite* (GOES) and the *Large Angle and Spectrometric Coronagraph* (LASCO), showed a significant correlation between flare parameters and CME kinetic energy (Yashiro and Gopalswamy, 2009). The just cited study also found that the most frequent flaring site was the center of the CME span.

As the eruptions are based on a storage and release mechanism of excessive magnetic free energy, photospheric magnetograms gave more insight into the forecasts than white light observations (Barnes et al., 2007). Thus, the studies associating flares with CMEs were extended to analyzing magnetic flux parameters of active regions (ARs), length of or strong magnetic gradient along the polarity inversion line (PIL), total photospheric free magnetic energy, etc. However, there seemed to be no individual feature that effectively discriminates between the flare alone and the flare associated with the CME group (Leka and Barnes, 2007). The statistical studies showed that the contribution of combined magnetic field features outscores the individual contribution. Bobra et al. (2014) had put forward all the possible magnetic features and created a pipeline to extract data termed as Space Weather HMI Active Region Patches (SHARPs). SHARPs track each active region from the *Solar Dynamics Observatory/Helioseismic and Magnetic Image* (SDO/HMI) data and derive 18 features based on various characteristics of the active region magnetic field parameters. The subset of

features that are more effective in prediction is not clearly defined, and no particular feature has shown a strong class separability, thus leaving space for improvement.

Machine learning (ML) and deep learning (DL) are good at analyzing underlying patterns of higher dimensional data. They have the capability to adapt and learn the characteristics of the existing data. Prediction of flare class, flare occurrence, CME occurrence (Yi et al., 2021; Zheng et al., 2021; Inceoglu et al., 2018; Abed, Qahwaji, and Abed, 2021; Park et al., 2018), solar energetic particles (SEPs) (Aminalragia-Giamini et al., 2021; Torres et al., 2022) have been studied using ML and DL. Interpretable machine learning has also been explored for SEP forecasting (Kasapis et al., 2022) using Space Weather MDI Active Region Patches (SMARPs) (Bobra et al., 2021). In the case of flares with CMEs, understanding the combination of the magnetic field features is complex; thus, the ML/DL approach can be used for classification and prediction. Qahwaji et al. (2008) explored the ML algorithms Cascade Correlation Neural Networks (CCNN) and Support Vector Machines (SVM) to study the relationship between flares and the associated CMEs. The study used flare intensity, flare duration, and the time interval between the peak time of the flare and its start and end times as inputs and achieved classification accuracy upto 0.88. Bobra et al. (2014) and Bobra and Ilonidis (2016) used extracted physical features of SHARPs as input to the SVM model for the prediction of CMEs that are associated with flares. They further classified the features as extensive and intensive based on their dependence on AR size and found that the combination of intensive parameters that are independent of AR size has more relevant information on the prediction. A True Skill Score (TSS) of 0.8 ± 0.2 was achieved in this study. Liu et al. (2019) extended the prediction of CME classification with time series information by using the DL model Recurrent Neural Network (RNN). They were able to forecast the CME associated with the flare 36 hours in advance with a TSS of around 0.6. SHARP parameters are extensively used as input features for the classification or prediction of flares and CMEs (Liu et al., 2017; Florios et al., 2018; Wang et al., 2019; Chen et al., 2019; Wang et al., 2020; Zhang et al., 2022a; Aktukmak et al., 2022; Sun et al., 2022; Zhang et al., 2022b). The studies mentioned above have been done using either instantaneous data or time series data and a specified ML algorithm. An exhaustive analysis of the performance of different ML algorithms is explored in this article, and the following issues are addressed.

- We compare the performance of eight different machine-learning models for the classification of whether a flare will be associated with a CME or not.
- The occurrence of flares associated with CMEs is much lower than that of flares without CMEs. The dataset is heavily imbalanced as seen in other solar physics problems. To address this, the performance of ML models before and after addressing the class imbalance through different techniques such as undersampling, oversampling, synthetic minority oversampling technique (SMOTE), and cost weight are explored.
- The best time lag to predict, best sampling methods, and best model are studied. We consider these criteria to study the performance of the models on a new dataset that includes temporal information.
- The performance of the ML models is studied using the wrapper method and the permutation feature importance method to understand more about the features significant for the prediction.

The article is organized as follows. Section 2 summarizes the input data source. Section 3 discusses the class imbalance and training methodology. Section 4 discusses the results with and without temporal information and analyzes the relevant features of the particular ML model using the wrapper and permutation feature importance methods. Section 5 gives a summary of the work and the scope for future work.
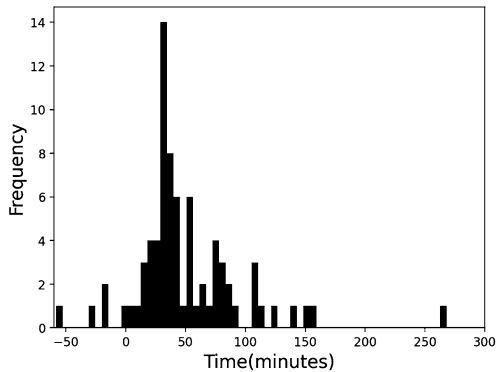
2. Data Source

The *Helioseismic and Magnetic Imager* on board the *Solar Dynamics Observatory* (Pesnell, Thompson, and Chamberlin, 2012; Schou et al., 2012) provides full-disk photospheric vector magnetic field data. Active regions (ARs) from HMI data are automatically identified and tracked. Some of the physical parameters are derived from AR patches known as Space Weather HMI Active Region Patches (SHARPs) (Bobra et al., 2014). The time resolution of the data is 12 minutes. The dataset is downloaded and processed following Bobra and Ilonidis (2016). Below are the brief steps that have been carried out in extracting the data.

1. The data on flares that occurred accompanied by CMEs are downloaded from the Database of Notifications, Knowledge, and Information (DONKI) website.[1]
2. The peak time of the flare is used to identify flare data from the GOES database and compare it with the DONKI database. If there is any missing/wrong AR number for the event in DONKI, the respective AR number is assigned from the GOES list.
3. The Joint Science Operations Center (JSOC) HMI database is queried to match the HARP number to a respective AR number.
4. Among a series of following time lags (8 h, 12 h, 24 h, 36 h, or 48 h), one is set as a time delay variable to extract the SHARP parameters before the peak time of the flare. These parameters are extracted at a single point in time (instantaneous) and not for the entire time series. We compare the performance of the models for the different time lags (8 h, 12 h, 24 h, 36 h, or 48 h) and choose the best time lag for the prediction, as described later in the article.

The article is focused on whether the SHARP parameters taken before the nth hour will be able to predict one between the two classes, i.e. flare only or flare with CME. The source code of the model is available online via github (https://github.com/Hemapriya-iit/Flare_classification_CMEs). The time difference distribution between the flare and the associated CME is shown in Figure 1. The histogram shows that most of the CMEs occur within $30-50$ minutes of the flare occurrence.

For the positive class, meaning flares are associated with CMEs, there were initially 161 data points that were downloaded from DONKI. Later, the data points where the AR number was not available and not matching a NOAA number were discarded. Further, based on the constraints such as: the orbital velocity of the spacecraft should be less than 3500 km s^{-1},

---

[1]https://kauai.ccmc.gsfc.nasa.gov/DONKI/search/.

high quality, and the event should fall within 70° of the central meridian, a few more data points were discarded. Finally, there are 85 data points that we use as being of a positive data class. For the negative class, meaning only a flare occurs, initially, there were 646 data points, while after the previously mentioned operations, there are finally 402 data points.

SHARP parameters are the features that give us a glimpse of the characteristics of active regions derived from vector magnetic fields obtained with HMI. These parameters can be a physical quantity, such as the total unsigned flux over the AR area, or proxy parameters, such as the total free magnetic energy. We use 18 such SHARP parameters for the period of 2011 – 2017 as input data, which are shown in Table 1. The twist parameter characterizes the degree to which the photospheric magnetic field deviates from the potential field. It is defined by the ratio between the vertical current density and the vertical magnetic flux density. The total unsigned magnetic flux in an AR, the sum of the unsigned magnetic flux (R) near the PIL, denotes the strength of the magnetic field. The free magnetic energy denotes the square of the difference between the potential magnetic field and the original magnetic field. The shear angle defines the angle between the potential transverse field and the observed transverse field. When the shear angle is large, it indicates that more free energy is stored and susceptible to flaring. The mean current helicity is the product between the vertical magnetic field density and the vertical electric current that can tell us how much magnetic energy is built up due to the twist and shear of the magnetic field. The mean angle of the field from the radial, the length of the PIL, the average gradient of the horizontal magnetic field, and the total unsigned net current per polarity can serve as a proxy for the energy stored in the current-carrying magnetic field and for how much net current is injected into the corona.

3. Methodology

Features are grouped based on their characteristics similar to Leka and Barnes (2007) and shown in Table 1. SHARP parameters taken at a single instance of time with different time lags (8 h, 12 h, 24 h, 36 h, and 48 h) before the flare occurrence are investigated. We study the performance of eight machine-learning models, namely Support Vector Machine (SVM), Logistic Regression (LR), Linear Discriminant Analysis (LDA), Decision Tree (DT), Random Forest (RF), Adaboost, Gradient Boost, and Extreme Gradient Boosting (XGBoost), in terms of effective prediction of the class that identifies the occurrence of CMEs along with flares. The model descriptions are given in Appendix A.

We have applied three steps in our work:

- (a) Address the class imbalance in SVM and apply the best technique to all models.
- (b) Develop a new dataset that will incorporate temporal information to the problem. This is done by taking the time difference of each feature from its preceding with 12 h time lag.
- (c) Apply the selected class-balancing approach to the new temporal dataset and study their performance.

3.1. Class Balancing Techniques

Class imbalance denotes when one class has a higher representation than the other class, which can become a major concern while developing ML models to predict a minority class. Here we are more interested in predicting the class of flares that are accompanied by CMEs, which are occasional events, as can be seen in Figure 2. Among the dataset of around 500 data points in the period 2011 – 2017, we observe that the number of samples of the class

Table 1 SHARP parameters.

Group	Parameter	Keyword	Formula
Horizontal gradients of the magnetic field	MEANGBH	Mean gradient of the horizontal field	$\|\nabla \bar{B}_h\| = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B_h}{\partial x}\right)^2 + \left(\frac{\partial B_h}{\partial y}\right)^2}$
	MEANGBT	Mean gradient of the total field	$\|\nabla \bar{B}_{\text{tot}}\| = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}$
	MEANGBZ	Mean gradient of the vertical field	$\|\nabla \bar{B}_z\| = \frac{1}{N} \sum \sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}$
Distribution of the vertical current density	MEANJZD	Mean vertical current density	$\bar{J}_z \propto \frac{1}{N} \sum \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right)$
	SAVNCPP	Sum of the modulus of the net current per polarity	$J_{z_{\text{sum}}} \propto \left\| \sum^{B_z^+} J_z dA \right\| + \left\| \sum^{B_z^-} J_z dA \right\|$
	TOTUSJZ	Total unsigned vertical current	$J_{z_{\text{total}}} = \sum \|J_z\| dA$
	MEANJZH	Mean current helicity (Bz contribution)	$\bar{H}_c \propto \frac{1}{N} \sum B_z \cdot J_z$
Current helicity	TOTUSJH	Total unsigned current helicity	$H_{c_{\text{total}}} \propto \sum \|B_z \cdot J_z\|$
	ABSNJZH	Absolute value of the net current helicity	$H_{c_{\text{abs}}} \propto \left\| \sum B_z \cdot J_z \right\|$
Twist parameter	MEANALP	Mean characteristic twist	$\alpha_{\text{total}} \propto \frac{\sum J_z \cdot B_z}{\sum B_z^2}$
Inclination angle	MEANGAM	Mean angle of the field from radial	$\bar{\gamma} = \frac{1}{N} \sum \arctan\left(\frac{B_h}{B_z}\right)$
Distribution of shear angles	MEANSHR	Mean shear angle	$\bar{\Gamma} = \frac{1}{N} \sum \arccos\left(\frac{\mathbf{B}^{\text{Obs}} \cdot \mathbf{B}^{\text{Pot}}}{\|B^{\text{Obs}}\| \|B^{\text{Pot}}\|}\right)$
	SHRGT45	Fraction of area with shear $\geq 45°$	Area with Shear $\geq 45°$/Total Area
Photospheric excess of magnetic energy density	MEANPOT	Mean photospheric magnetic free energy	$\bar{\rho} \propto \frac{1}{N} \sum \left(\mathbf{B}^{\text{Obs}} - \mathbf{B}^{\text{Pot}}\right)^2$
	TOTPOT	Total photospheric magnetic free energy density	$\rho_{\text{tot}} \propto \sum \left(\mathbf{B}^{\text{Obs}} - \mathbf{B}^{\text{Pot}}\right)^2 dA$
Magnetic field parameters	USFLUX	Total unsigned flux	$\Phi = \sum \|B_z\| dA$
	AREA_ACR	Area of strong field pixels in the active region	Area $= \sum$ Pixels
	R_VALUE	Sum of flux near the PIL	$\Phi = \sum \|B_{\text{LOS}}\| dA$ within R mask

flares with CMEs is around 80 while the flare-only class consists of around 420 samples. Feeding the imbalanced dataset to ML models leads to a bias towards the majority class. To overcome it, we have tried to address the class imbalance issues through various techniques such as undersampling majority class, oversampling minority class, and Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). We have compared the performance of the SVM model with respect to each sampling technique and the class-weight method (Bobra and Ilonidis, 2016). These techniques are described briefly here.

- Random Under Sampling (RUS) drops random data points from the majority class, thus balancing the number of data belonging to each class. This method works well when we have a large dataset, as the characteristics of the majority class will not change drastically

Figure 2 Representation of the number of events in each class for the period 2011–2017.
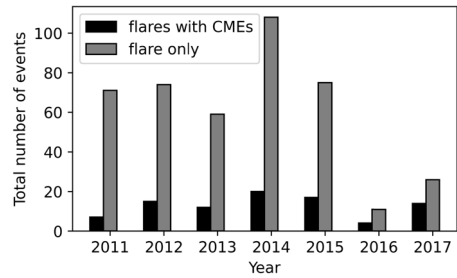
after dropping a few data oints. It could be a disadvantage for a small dataset, as the model may not learn the characteristics of the majority class well.

- Random Over Sampling (ROS) randomly duplicates the data points of the minority class. For small dataset, this method will help in achieving a balanced dataset. These methods are naive resampling of the existing data without assuming much about the data distribution.
- Replicating the minority class data make the decision region more specific. It can lead to reduced generalization, thus overfitting. In such cases, to overcome the drawback, synthetic data can be generated after certain operations on the original minority class. Based on data points that are close together in the feature space, the SMOTE method creates synthetic minority samples. The Euclidean distance is used to determine the k nearest neighbors for a minority class sample, let us say n. The distance between the n vector and any or all of the k nearest neighbors is calculated and multiplied by a random value between (0,1). The resultant is added with the feature vector n to generate a synthetic sample. In this work, $k = 5$ has been used and implemented using the Scikit-Learn module sklearn.neighbors.NearestNeighbors.
- Rather than increasing or decreasing the number of data points, the performance of the model for the minority class can be increased by introducing the class weights. Bobra and Ilonidis (2016) have used the class-weight parameter to give more weights in the ratio of 1:6.5 to the minority class. This will lead to a higher penalty by the loss function for any misclassification of the minority class.

We have used the above-mentioned sampling techniques to address the class imbalance issue and have compared the effectiveness of each sampling technique in the performance of the model for the present problem.

3.2. Training Methodology

The data have been initially standardized with median and standard deviation similar to Bobra and Ilonidis (2016). The class labels '1' and '0' are assigned to the positive class (flares with CMEs) and negative class (flares without CME), respectively. The training of the ML models is carried out as follows. We split the dataset randomly as training and testing using the stratified k-fold method. The stratified k-fold method ensures an equal proportion of positive and negative classes in the training and testing dataset. In this method, the dataset will be divided into k folds, among which $(k-1)$ sets will be used for training, and the remaining fold will be used as a test dataset until all the folds are exhausted for testing. Here, it is to be noted that only the training set is balanced using class balancing techniques. This balanced dataset is then fed to the eight machine-learning algorithms for training. Each trained model is then evaluated on the class imbalanced test dataset. The same procedure
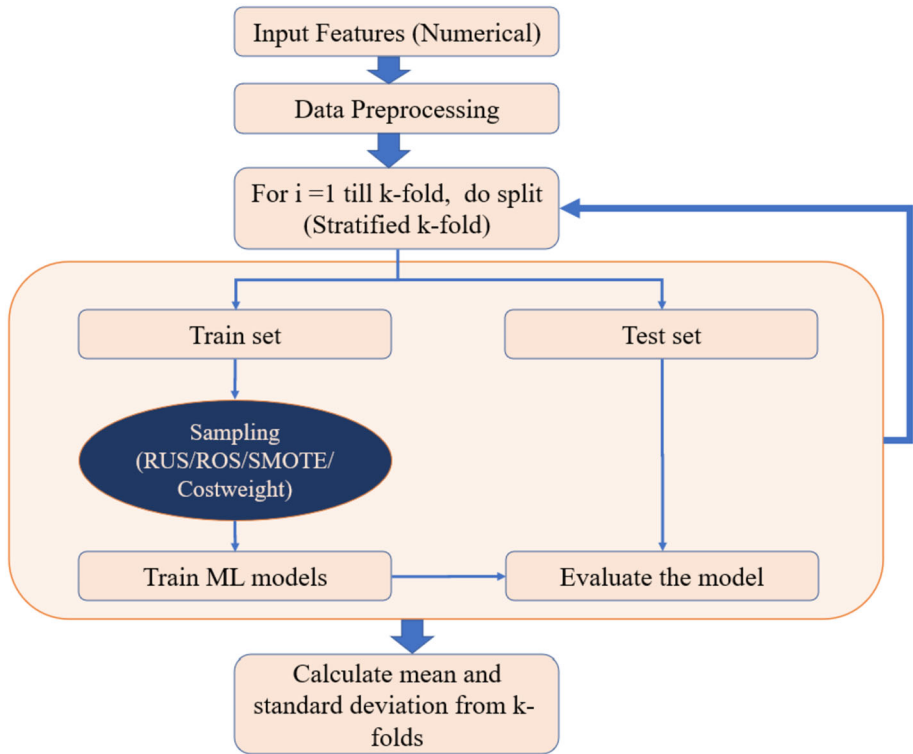
Figure 3   Training process of eight machine-learning models with different class-balancing techniques.

is repeated for each *k* fold iteratively, and the evaluation metrics is calculated. Figure 3 represents the flow of the training and testing that has been carried out in this article. Finally, the mean and the standard deviation of the performance metrics are calculated to understand the capability of the generalization of the model and the confidence bound. Here the model is evaluated for six folds.

As this is a binary classification problem, the models are evaluated using the confusion matrix. The true skill score (Woodcock, 1976), probability of detection, and false alarm rate are derived from the confusion matrix and used as the indicator for the model performance. The description of the metrics is given in Appendix C.

4. Results and Discussion

The performances for different ML algorithms are evaluated for different time lags, i.e. 8 h, 12 h, 24 h, 36 h, and 48 h before the event occurrence similar to Ahmed et al. (2013), and Bobra and Ilonidis (2016). The class imbalance is addressed and studied through different sampling techniques. Later the best-performing methodology and time lag are adopted and studied using the dataset with temporal information.
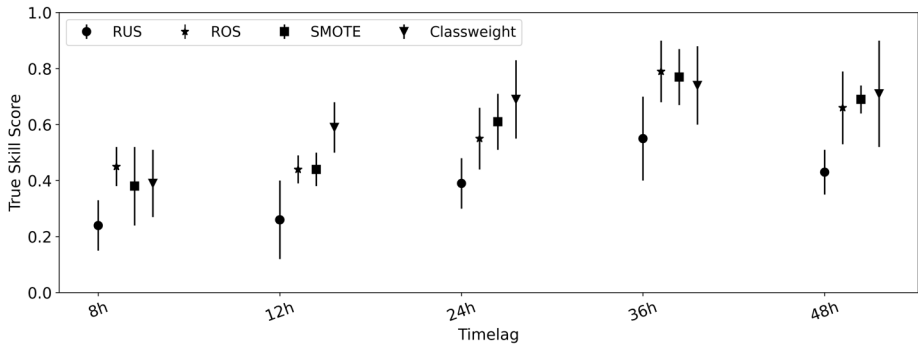
Figure 4 Performance of SVM for different time lags and class-balancing methods.
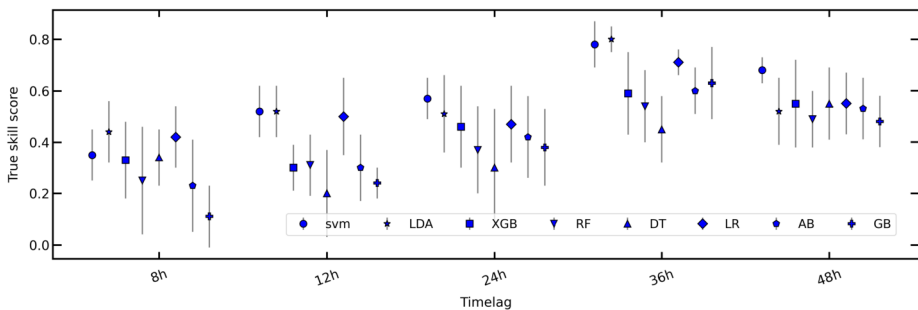

Figure 5 Performance of all ML models for different time lags for the SMOTE method.

4.1. Comparison of the Performance of Different ML Algorithms

Bobra and Ilonidis (2016) have shown that the ML model SVM was more effective in classifying between flares associated with CMEs and flare-alone events. They addressed the class imbalance of the dataset using a class-weight approach. Hence, we first experiment with different class balance approaches on the SVM model to study their impact at different time lags. Figure 4 summarizes the performance of the SVM model evaluated in terms of the True Skill Score (TSS) metric. The class balancing approach random undersampling method is found not suitable as the dataset is small, while class weight, ROS, and SMOTE show improved performance in the TSS. Though ROS is found to balance the class distribution, it does not add any additional information to the model. Hence, we prefer SMOTE for class balancing, as it generates new samples drawn from the near feature space. SMOTE class balancing is applied to all models for different time lags, and the results are shown in Figure 5. We observe that the 36 h time lag results are good compared to other time lags for all models. The time lag of 36 h and SMOTE technique are fixed to study the results from now on in this article. Figure 6 compares the performance of eight machine-learning models for balanced and imbalanced datasets using various class-balancing approaches for different time lags. It can be seen that the impact of class balancing techniques is negligible for LDA. This can be because both classes are assumed to share the same covariance matrix.

Figure 7 shows the performance of all models for a 36 h time lag. Overall we find that the SVM, LDA, and Logistic Regression perform well among the eight models we evaluate.
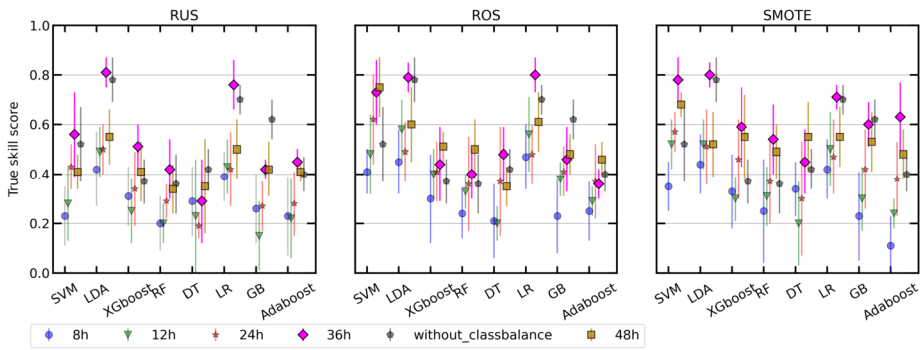
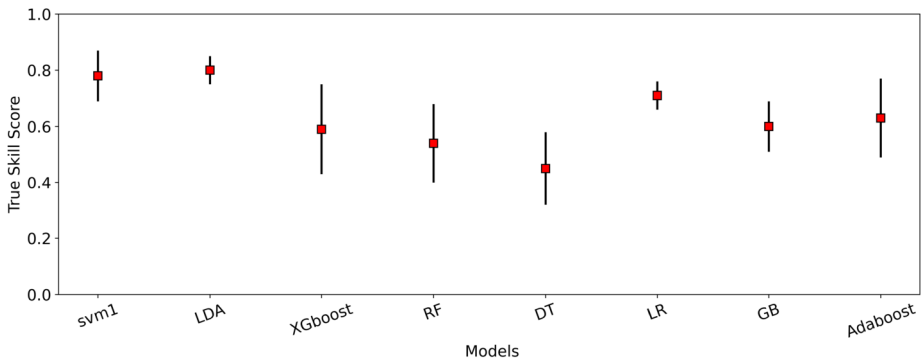**Figure 6** Performance of all ML models for different time lags for different class-balancing techniques.

Figure 7 The performance of all ML models with uncertainty for 36 h time lag and SMOTE class-balancing method.

Table 2 Performance metrics for eight machine-learning models.

Model	TSS	FAR	FDR	PoD
SVM	**0.78 ± 0.09**	**0.02 ± 0.02**	**0.12 ± 0.1**	**0.8 ± 0.08**
LDA	**0.8 ± 0.05**	**0.09 ± 0.02**	**0.31 ± 0.05**	**0.88 ± 0.05**
XGBOOST	0.59 ± 0.16	0.06 ± 0.02	0.3 ± 0.11	0.65 ± 0.17
RF	0.54 ± 0.14	0.04 ± 0.04	0.23 ± 0.13	0.59 ± 0.17
DT	0.45 ± 0.13	0.14 ± 0.04	0.52 ± 0.05	0.59 ± 0.15
LR	0.71 ± 0.05	0.1 ± 0.03	0.37 ± 0.08	0.81 ± 0.03
GB	0.6 ± 0.09	0.15 ± 0.04	0.48 ± 0.07	0.75 ± 0.07
AdaBoost	0.63 ± 0.14	0.06 ± 0.02	0.27 ± 0.08	0.68 ± 0.16

LDA shows a TSS score of 0.8 ± 0.05, and SVM of 0.78 ± 0.09. Class balancing through SMOTE increases the performance of SVM from 0.52 using an unbalanced dataset to 0.8. Other metric comparisons of all the models are shown in Table 2. We find that a false alarm rate is less in SVM than in LDA, and the Probability of detecting a CME is higher in LDA, around 88%, than in SVM, around 80%.
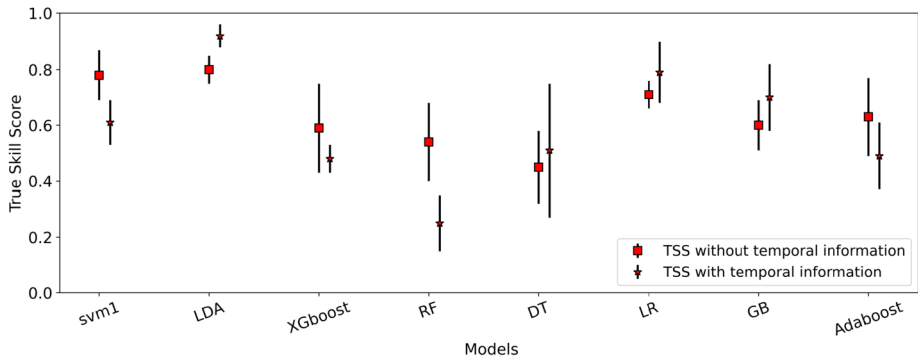
Figure 8 Overall performance comparison of all models with and without temporal information.

Till now, 36 h time-lagged data are used as independent parameters without incorporating the previous time lag information. We have seen that the 36 h time lag performs well in the previous method. Hence, a simple experiment is carried out here incorporating the additional information of difference with its preceding 12 h time lag. This is done to see if the resultant model benefits from the additional information on what has changed from the previous 12 hours. LDA shows improved performance from 0.8 to 0.9 in the TSS score. The results of all models for both datasets are shown in Figure 8. The results demonstrate that the performance of the remaining model is degraded except for LDA, LR, GB, and DT. We try to tune the hyperparameters of SVM, but even so, the performance of SVM degrades from 0.8 to 0.6, while LDA shows an improvement from 0.8 to 0.9. All the evaluation matrices are summarized in Table 3. FAR increases for SVM, while it decreases for LDA. LDA shows an increased performance of TSS score, up to 0.92 ± 0.04. LDA detects the occurrence of a CME well on an average of 96% of the time, with a lower FAR of 4%. The SVM performance falls down with a maximum TSS score of 0.61 ± 0.08 with FAR of 9%. We also observe that the Logistic Regression model performs better with a TSS score of 0.79 ± 0.11 and detects the occurrence of CMEs 85% of the time.

As the testing set is independent, the data is shuffled randomly for k-fold cross-validation. There might be cases where the different flare cases from the same AR have been shared between the training and testing set. Thus, we try assigning a whole year as an independent dataset and the remaining years for training from $2011 - 2018$ and evaluate the models SVM and LDA with and without temporal information. The results are shown in Table 4. We compare the performance of SVM and LDA with and without temporal information along with the results from the class weight that was carried out in Bobra and Ilonidis (2016) to address the class imbalance. The number of positive and negative classes in training and testing are also mentioned in Table 4. We observe that for SVM and LDA, the testing performance varies for different years. This can be due to the characteristics of the phase of the solar cycle and the varying size of the test dataset.

4.2. Model Interpretation

We make an attempt to comprehend the features that have influenced the prediction of the model in order to understand its role. Before running any ML model through training, we use the feature importance technique K-best to assess the contributions of each individual to the class prediction. After training, we select the SVM and LDA algorithms that did well

Table 3 Metrics for eight ML models for data with temporal information.

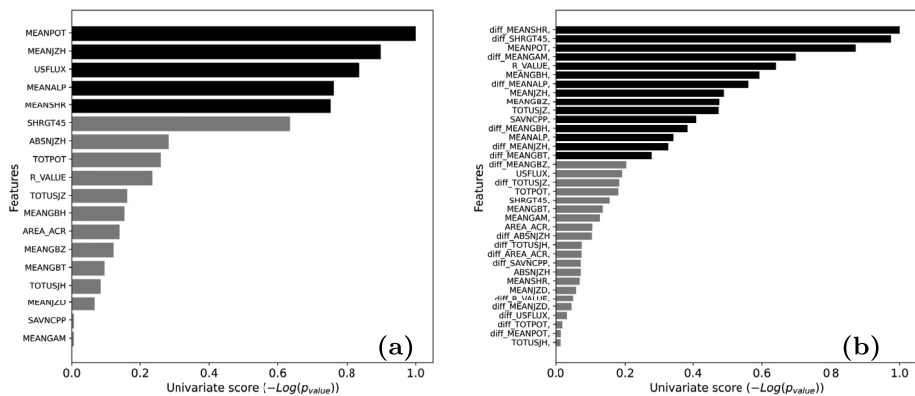| Model	TSS	FAR	FDR	PoD
SVM	0.61 ± 0.08	0.09 ± 0.02	0.41 ± 0.06	0.7 ± 0.08
LDA	$\mathbf{0.92 \pm 0.04}$	$\mathbf{0.04 \pm 0.03}$	$\mathbf{0.16 \pm 0.11}$	$\mathbf{0.96 \pm 0.04}$
XGBOOST	0.48 ± 0.05	0.1 ± 0.02	0.49 ± 0.06	0.58 ± 0.05
RF	0.25 ± 0.1	0.06 ± 0.02	0.5 ± 0.1	0.3 ± 0.09
DT	0.51 ± 0.24	0.15 ± 0.04	$.55 \pm .13$	$.66 \pm 0.2$
LR	0.79 ± 0.11	0.05 ± 0.03	0.23 ± 0.12	0.85 ± 0.11
GB	0.7 ± 0.12	0.07 ± 0.04	0.29 ± 0.12	0.77 ± 0.14
AdaBoost	0.49 ± 0.12	0.09 ± 0.01	0.47 ± 0.06	0.59 ± 0.11

Figure 9 Feature selection results using the K-best method for data with (**a**) without temporal information and (**b**) with temporal information.

in both methods with and without temporal information. We use the wrapper technique to investigate which feature subsets are responsible for good prediction. Later, we intend to get an explanation of the model by evaluating both models with the permutation feature importance method to determine the top few features responsible for the good prediction of the model.

4.2.1. Feature Importance by Ranking

The contributions of individual features for the predictions are evaluated using a ranking technique called Select K-best method. This method uses the analysis of variance (ANOVA) between two different populations. This method ranks the relationship of each individual feature with respect to the target. The p-values returned are used to calculate the significance of the ranked features for the prediction. We calculate both data with and without temporal information after inspecting the best time lag, i.e. 36 h. The features that are important when using the 36 h time lag are shown in Figure 9. We observe that the total photospheric excess free energy (MEANPOT), current helicity (MEANJZH), total unsigned flux (USFLUX), twist parameters(MEANALP), and shear parameters (MEANSHR, SHRGT45) play an important role in associating the flares with CMEs. Adding the temporal information to the

Table 4 TSS metrics for SVM and LDA with one independent year test data.

Test year	Train year	SVM	LDA	SVM temporal	LDA temporal
***k*fold**		**0.78 ± 0.09**	**0.8 ± 0.05**	**0.61 ± 0.08**	**0.92 ± 0.04**
2011	2012, 2013, 2014, 2015, 2016, 2017	0.74	0.93	0.23	0.50
$test\_n = 63$	$train\_n = 339$				
$test\_p = 7$	$train\_p = 78$				
2012	2011, 2013, 2014, 2015, 2016, 2017	0.54	0.63	0.63	0.55
$test\_n = 73$	$train\_n = 329$				
$test\_p = 15$	$train\_p = 70$				
2013	2011, 2012, 2014, 2015, 2016, 2017	0.62	0.87	0.73	0.82
$test\_n = 59$	$train\_n = 343$				
$test\_p = 11$	$train\_p = 74$				
2014	2011, 2012, 2013, 2015, 2016, 2017	0.33	0.67	0.28	0.33
$test\_n = 105$	$train\_n = 297$				
$test\_p = 19$	$train\_p = 66$				
2015	2011, 2012, 2013, 2014, 2016, 2017	0.58	0.92	0.59	0.65
$test\_n = 74$	$train\_n = 328$				
$test\_p = 17$	$train\_p = 68$				
2016	2011, 2012, 2013, 2014, 2015, 2017	0.50	0.87	1	1.0
$test\_n = 8$	$train\_n = 394$				
$test\_p = 3$	$train\_p = 82$				
2017	2011, 2012, 2013, 2014, 2015, 2016	0.40	0.79	0.14	0.17
$test\_n = 20$	$train\_n = 382$				
$test\_p = 13$	$train\_p = 72$				

data, we perform the same feature selection method. We can see in Figure 9b that the temporal information of shear parameters (MEANSHR, SHRGT45), distribution of inclination angle parameter (MEANGAM), and twist parameters (MEANALP) are significant in the prediction when using a $24 - 36$ h time lag along with total photospheric excess free energy (MEANPOT) and the magnetic field near the PIL (R_value).

4.2.2. Wrapper Method

In this method, the model performance of the SVM and LDA is evaluated by adding or removing each feature from the existing feature subsets, beginning with the null set. This method is known as bidirectional wrapping, a way of feature selection method after the model is trained. If the newly added feature degrades the performance of the model, either
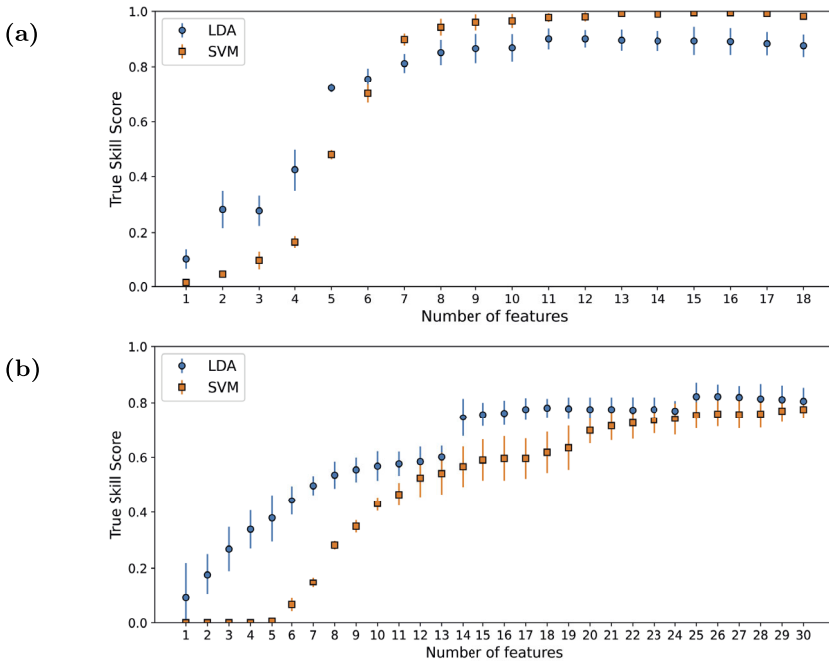
Figure 10 Performance of the model with a combination of features using the wrapper method for data (**a**) without and (**b**) with temporal information.

that feature or other newly added features that degrade performance will be removed, and the next feature will be examined. In this approach, the wrapper method examines the performance of the model by simultaneously adding features that improve performance while discarding those that degrade performance. The performance of the model is shown in Figure 10a-b for data without and with temporal information, respectively. The results shown are with inbuilt k-folds in the wrapper model where both the training and test set are balanced using SMOTE. We study the feature sets, which have contributed to the increase in the performance of the model for both SVM and LDA.

- For data without adding temporal information, the wrapper method results are seen in Figure 10a. The metrics mentioned here are for the training dataset. A detailed combination of the features is given in Appendix B. The performance of the SVM model increases from 0.48 and 0.70 to 0.94 when the features related to the inclination angle of the horizontal field (MEANGAM), the shear angle (MEANSHR), and the magnetic field near PIL (R_value) are added. The performance of the model reaches its peak around the first seven features, with a drastic jump from the subsequent number of features; later, it saturates. The top eight features are related to the magnetic field parameters (USFLUX, AREA_ACR, R_VALUE), the current helicity (MEANJZH, ABSNJZH), the photospheric excess magnetic free energy (MEANPOT, TOTPOT), the inclination angle of the horizontal field (MEANGAM), and the shear angle (MEANSHR). For LDA, we observe that there is a sharp jump in the performance from 0.42 to 0.75 from the 4th to 6th feature when the inclination angle parameter (MEANGAM) and the shear parameters (MEANSHR) are added to the feature subset. The top eight features that are observed in LDA are USFLUX, MEANGBH, MEANGAM, MEANGBZ, TOTUSJZ, MEANSHR,
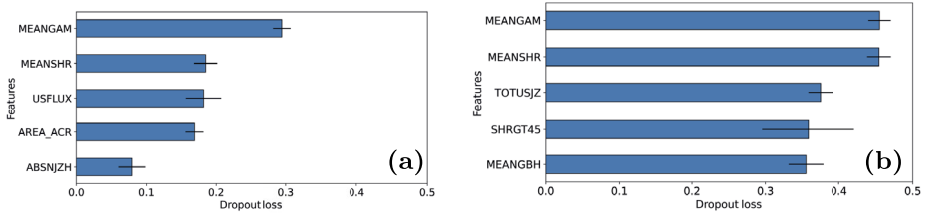
Figure 11 Interpretations for model (**a**) SVM and (**b**) LDA showing first few important features with greater dropout loss for data without temporal information.

AREA_ACR, and ABSNJZH in no particular order. Thus, we can see that shear and inclination angle parameters have played important roles in improving the performance of both models.

The wrapper method is also applied to the data that includes temporal information to the existing features, and the results are analyzed.

- From Figure 10b, for the SVM model, we observe that the performance saturates around the 11th and 19th features. For the first few features, the model shows no significant performance. The features are USFLUX, MEANGBT, MEANJZH, and MEANPOT. After temporal information of the shear parameters (MEANSHR) and total photospheric excess free energy (MEANPOT) are added, the model shows improvement in performance. Apart from these features, the top features are USFLUX, MEANGBT, MEANJZH, MEANPOT, MEANGBH, SAVNCPP, TOTPOT, R_VALUE, temporal information of MEANSHR, MEANPOT, and AREA_ACR. Here the inclination angle parameters are not observed in the top few features. For LDA in Figure 10b, the performance saturates after 13 features. Some of the important features that contribute to the performance are temporal information of MEANGBT, MEANPOT, SHRGT45, TOTUSJH, MEANGBH, MEANGBZ, TOTUSJZ, and MEANSHR. Similarly, adding MEANGAM, MEANPOT, SHRGT45, and TOTPOT has improved the performance of the model.

4.2.3. Permutation Feature Importance

The permutation feature importance method is used to understand the model interpretation rather than the individual predictions. After training a given model on the data, the performance of the model is examined by removing the specific feature and computing the dropout loss of the model. The larger the dropout loss, the more important the feature is, for prediction. We investigate SVM and LDA for data with and without temporal information. For data without adding temporal information, Figure 11a of the SVM model suggests that MEANGAM, USFLUX, AREA ACR, MEANSHR, and ABSNJZH are some of the top features that contribute to the predictions of the model, with the inclination angle parameter MEANGAM having the highest dropout loss. The order may vary slightly for different seeds, but not considerably. The dropout loss for LDA is larger for the shear parameter MEANSHR and the inclination angle parameter MEANGAM, as shown in Figure 11b. Figure 11a-b shows the performance of all k-folds within the uncertainty, and we observe that the classification is majorly based on the amount of magnetic flux present, the extent of the sheared magnetic field, the current helicity, and the inclination angle of the horizontal field.

For data that includes temporal information, we observe that LDA performs the best when compared to SVM, as mentioned in Table 3. In Figure 12a, MEANGBT, MEANPOT,
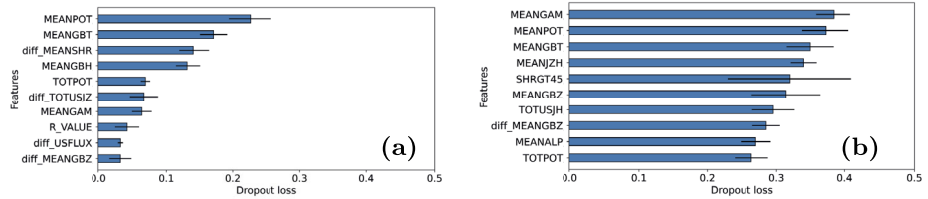
Figure 12   Interpretations for model (**a**) SVM and (**b**) LDA showing the top few features with their dropout loss for data including temporal information.

MEANGBH, and TOTPOT show more importance for the SVM model considering all folds, which seems to vary significantly from the data without temporal information. We suppose that emphasizing the features related to the gradient of the magnetic field and magnetic free energy more than the shearing or inclination parameters can be one of the reasons for the degradation in the performance of SVM. In LDA, the features related to the gradient of the magnetic field features (MEANGBT, MEANGBZ), photospheric excess free energy (MEANPOT), and temporal information of the gradient of the magnetic field have been added as significant features, as shown in Figure 12b. However, LDA retains the inclination angle parameter (MEANGAM) as one of the top features. Thus, we wish to highlight here that the feature importance varies from model to model, and also which can be used to understand the reason for the deteriorating performance of the specific model.

5. Summary

The evaluation of the performance of eight ML models on the prediction of CMEs that occur with flares is carried out in this article. We attempt to address the prevalent class imbalance encountered in the majority of the solar physics problems where we are interested in occasional occurrences of the events. We systematically explore the class balancing techniques ROS, RUS, and SMOTE for this task of predicting CMEs associated with flares by analyzing it on the already well-performing SVM model. Later, the best-performing approach SMOTE is applied to all eight ML models using different time lags. The parameters taken before the 36 h time lag before the flare is found to be the best for the predictions. We find that both SVM and LDA perform well, with a TSS score of around 0.8. We find that while SVM is affected by class imbalance and thus needs some class balancing techniques, LDA can also be efficient with an imbalance dataset, making LDA an efficient model in the prediction. Additionally, the temporal information of the feature with its preceding time lag is added to the existing dataset to study their contributions. We observe that LDA increases its performance with a TSS score up to 0.9, whereas SVM is seen deteriorating in its performance. This kind of random shuffling implies the possibility of an AR getting shared between training and testing sets, though it produces different events. Hence, we also validate the model through year-wise testing data. Both LDA and SVM show varying TSS performance for different years of testing. The size of the testing set and the solar cycle characteristics over different years may have contributed to this performance variance and thus need to be carefully considered while developing ML models. A detailed analysis in this respect will be carried out in our future work.

The relevance of the features in the prediction is studied using feature importance techniques. The K-best method is selected to study the contribution of individual features with

respect to the target in CME prediction. Certain combination of features performs better than the individual feature contributions in predicting the CME occurrence with flares. Thus, the wrapper method is applied after training the model, and the feature subsets are studied to understand the top few combinations that have contributed significantly to the prediction. We observe that the inclination angle parameter MEANGAM, shear angle parameters, and magnetic field parameters significantly improve the performance of both LDA and SVM. We further attempt to study the model predictions using permutation-based variable importance measures. We observe that the inclination angle parameter MEANGAM, shear parameters, and current helicity parameters show more dropout loss in the performance of the model when such features are permuted. The SVM model shows a decrease in performance when the temporal information is added, as we observe that it gives more importance to the gradient of the magnetic field information than to the inclination angle parameters, which might not be much significant to make a decision on the occurrence of a CME. Thus, the interpretation of the models can also help us understand the reason behind their poor performance. In the generation of synthetic samples to address class imbalance, more advanced models such as generative models, i.e. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which can reproduce similar data for the minority class will be carried out in future work.

Appendix A: Machine Learning Models

- Linear models
  - Support Vector Machine (SVM) (Cortes and Vapnik, 1995) can be used effectively for classification and regression. In a high-dimensional space, SVM sets a hyperplane for binary classification. This is achieved by transforming a nonlinear space to a linear space using a kernel at the maximized distance between the classes from the nearest training data points. As the margin between two classes becomes larger, SVM classifies effectively in an ideal scenario. However, in a real case, there will always be a few cases on the other side that cannot be separable, so support vectors can be there on either side of the class. Finding an optimized method between minimizing the support vectors and maximizing the separable margin allows defining a good model leading to a reduced misclassification rate. Support vector machine can handle both linear and nonlinear data using kernels. In this article, we use a radial basis function (RBF) kernel similar to Bobra and Ilonidis (2016).
  - Logistic regression is a statistical method that is used to estimate the probability of a data point belonging to either class. This is achieved using a sigmoid function in which, above a particular threshold, the data point belongs to one class, and below the threshold, it belongs to the other class.
- Discriminant model
  - Linear discriminant analysis (LDA) (Tharwat et al., 2017) assumes that the input data is Gaussian. For each class, LDA estimates the mean and variance from the given data. During prediction, LDA tries to get the probability of the new input belonging to a particular distribution of a class using the Bayes theorem. The class which has a higher probability is assigned as the respective prediction for the input. It further acts as a dimensionality reduction technique by projecting features from a higher-dimension to a lower-dimension space.
- Decision based models

– A decision tree classifies data according to the attribute associated with the root node and later moves down to the leaf node by making certain decisions. Decision trees are naturally immune to a correlation between variables and resistant to outliers. However, decision trees are prone to overfitting, producing biased trees based on a particular majority class.

• Ensembling models
  – As mentioned above, decision trees are a bit unstable and suffer from generalization issues. Ensemble techniques in the form of bootstrapping or bagging that draw several different decisions from the dataset can be helpful in such cases. The prediction, thereby using majority voting or average outcomes, makes the structure more stable, called as Random Forest (RF) (Breiman, 2001). RF is more stable, accurate, and adaptable to both classification and regression. Random Forest helps in generalization and reduces overfitting by changing the algorithms that are learned by subtrees each time. Random Forest may suffer the worst performance if there are more correlated variables.
  – Boosting (Freund, 1995) is a sequential process where instead of averaging the outcomes, the decisions are weighted and adjusted in sequence models according to the errors obtained from the previous model. The Adaboost model is an ensembling model of weak learners where each model is boosted by the performance of the ensembles to obtain a good classifier with good accuracy.
  – Extreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016) is an ensembling algorithm enclosing groups of individual learners; it uses gradient descent to minimize the loss function. Further modeling is designed using a regularization parameter and inbuilt cross validation; thus, reducing overfitting and generalization error. Tuning and visualization of Xgboost is a bit complex compared to Adaboost and Random Forest.

Appendix B: Wrapper Features

In Table 5 and 6, the first column represents the number of features followed by the performance metrics and the features that contributed to the performance of SVM. The features that contribute to the performance of LDA are shown in Table 6.

Appendix C: Metrics

A True Positive (TP) case is defined here as when both the actual and predicted classes are eruptive flares. A False Positive (FP) is when the actual is a confined flare, whereas the predicted class is an eruptive flare, thus leading to false alarms. When the actual event is an eruptive flare, whereas the predicted is falsely classified as a confined flare, it is termed as False Negative (FN). True Negative (TN) refers to when both actual and prediction classes are confined flares. Thus, from the confusion matrix, it is understood that a good model should have fewer False Positives and also False Negatives. Performance metrics should be carefully chosen before assessing the performance of the model with class imbalance. Metrics such as accuracy show higher performance due to better classification of the majority class than the minority class. Since we have more confined flare events, using accuracy as a metric will lead to a biased model evaluation. Hence, we evaluate a few other performance metrics based on TP, FP, TN, and FN obtained from the confusion matrix. True Skill Score, Probability of Detection (PoD), False Alarm Rates (FAR), and False Detection Rate (FDR) are a few metrics that are described in detail below.

Table 5 Wrapper features for SVM.

No.	TSS	Feature subsets
1	0.014	['MEANJZH']
2	0.044	['MEANJZH', 'ABSNJZH']
3	0.094	['MEANJZH', 'TOTPOT', 'ABSNJZH']
4	0.166	['MEANJZH', 'TOTPOT', 'AREA_ACR', 'ABSNJZH']
5	0.481	['MEANJZH', 'MEANPOT', 'TOTPOT', 'AREA_ACR', 'ABSNJZH']
6	0.702	['MEANJZH', 'MEANPOT', 'MEANGAM', 'TOTPOT', 'AREA_ACR', 'ABSNJZH']
7	0.898	['MEANJZH', 'MEANPOT', 'MEANGAM', 'TOTPOT', 'MEANSHR', 'AREA_ACR','ABSNJZH')]
8	0.942	['USFLUX', 'MEANJZH', 'MEANPOT', 'MEANGAM', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'ABSNJZH']
9	0.96	['USFLUX', 'MEANJZH', 'MEANPOT', 'MEANGAM', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH']
10	0.965	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'MEANGAM', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH')]
11	0.977	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'MEANGAM', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE']
12	0.98	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGAM', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE']
13	0.992	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANALP', 'MEANGAM', 'MEANJZD', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE']
14	0.99	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANALP', 'MEANGAM', 'MEANJZD', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH']
15	0.995	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANALP', 'MEANGAM', 'MEANJZD', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR','AREA_ACR', 'R_VALUE', 'ABSNJZH']
16	0.995	['USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANJZD', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH']
17	0.992	['USFLUX', 'MEANGBT', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANJZD', 'TOTUSJZ', 'SAVNCPP','TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH'],]
18	0.982	['USFLUX', 'MEANGBT', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'MEANJZD', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH')]

- Probability of Detection (PoD): The PoD is described as the probability that our model will detect the occurrence of eruptive flare events. Thus, we calculate the ratio between true positive events and total actual eruptive flares. It is also known as True Positive Rate (TPR). It can be described by Equation 1:

$$PoD = \frac{TP}{TP + FN}.$$ (1)

- False Alarm Rate (FAR): The performance of a model is considered good when there are fewer false alarms. Thus, minimizing the ratio of wrongly categorized eruptive flare events to the total number of actual confined flare events should be considered, which is

Table 6 Wrapper features for LDA.

No.	TSS	Feature subsets
1	0.09	[('ABSNJZH',)]
2	0.28	[('SAVNCPP', 'ABSNJZH')]
3	0.27	[('SAVNCPP', 'AREA_ACR', 'ABSNJZH')]
4	0.42	[('USFLUX', 'SAVNCPP', 'AREA_ACR', 'ABSNJZH')]
5	0.72	[('USFLUX', 'MEANGAM', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
6	0.75	[('USFLUX', 'MEANGBH', 'MEANGAM', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
7	0.85	[('USFLUX', 'MEANGBH', 'MEANGAM', 'MEANGBZ', 'MEANSHR', 'AREA_ACR','ABSNJZH')],
8	0.81	[('USFLUX', 'MEANGBH', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
9	0.86	[('USFLUX', 'MEANGBT', 'MEANGBH', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
10	0.86	[('USFLUX', 'MEANGBT', 'MEANGBH', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
11	0.9	[('MEANGBT', 'MEANPOT', 'SHRGT45', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
12	0.9	[('MEANGBT', 'MEANPOT', 'SHRGT45', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'SAVNCPP', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
13	0.89	[('MEANGBT', 'MEANPOT', 'SHRGT45', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
14	0.89	[('USFLUX', 'MEANGBT', 'MEANPOT', 'SHRGT45', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'ABSNJZH')],
15	0.89	[('USFLUX', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR','AREA_ACR', 'R_VALUE', 'ABSNJZH')],
16	0.89	[('USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT','MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH')],
17	0.88	[('USFLUX', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH', 'MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'MEANJZD', 'TOTUSJZ', 'SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR','R_VALUE', 'ABSNJZH')],
18	0.87	[('USFLUX', 'MEANGBT', 'MEANJZH', 'MEANPOT', 'SHRGT45', 'TOTUSJH','MEANGBH', 'MEANALP', 'MEANGAM', 'MEANGBZ', 'MEANJZD', 'TOTUSJZ','SAVNCPP', 'TOTPOT', 'MEANSHR', 'AREA_ACR', 'R_VALUE', 'ABSNJZH')]

denoted by False Alarm Rate (FAR), otherwise known as False Positive Rate (FPR). FAR is calculated by Equation 2:

$$FAR = \frac{FP}{FP + TN}. \tag{2}$$

The good model will maximize the performance of PoD, but there exists a trade-off while trying to minimize FAR; thus, an optimum balance should be found between both metrics.

- False Discovery Rate (FDR): FDR can be defined as the rate at which false alarms occur among all predicted eruptive flare events. FDR is given as in Equation 3:

$$FDR = \frac{FP}{FP + TP}.$$ (3)

- True Skill Score (TSS): In order to avoid bias towards the majority class events, we try to evaluate the model based on the metric TSS. TSS can be seen as the difference between PoD and FAR. Thus, a good model with a high TSS score will have more PoD of eruptive flare events while minimizing the FAR. However, maximizing the only metric TSS may lead to more False Positives. TSS ranges from -1 to $+1$, with $+1$ as the best measure for our desired class. TSS can be evaluated as in Equation 4. TSS is not affected by the class imbalance; hence, we have considered it as our base metric to evaluate our performance of the model.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN}.$$ (4)

Author contributions Hemapriya Raju developed the model and tested different approaches. Hemapriya Raju wrote the draft manuscript. Saurabh Das supervised the manuscript. Both Hemapriya Raju and Saurabh Das contributed to the final version of the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests The authors declare no competing interests.

References

Abed, A.K., Qahwaji, R., Abed, A.: 2021, The automated prediction of solar flares from sdo images using deep learning. *Adv. Space Res.* **67**(8), 2544.

Ahmed, O.W., Qahwaji, R., Colak, T., Higgins, P.A., Gallagher, P.T., Bloomfield, D.S.: 2013, Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Solar Phys.* **283**(1), 157. DOI.

Aktukmak, M., Sun, Z., Bobra, M., Gombosi, T., Manchester, W.B. IV, Chen, Y., Hero, A.: 2022, Incorporating polar field data for improved solar flare prediction. *Front. Astron. Space Sci.* 9. DOI.

Altschuler, M.D., Trotter, D.E., Orrall, F.Q.: 1972, Coronal holes. *Solar Phys.* **26**, 354.

Aminalragia-Giamini, S., Raptis, S., Anastasiadis, A., Tsigkanos, A., Sandberg, I., Papaioannou, A., Papadimitriou, C., Jiggens, P., Aran, A., Daglis, I.A.: 2021, Solar energetic particle event occurrence prediction using solar flare soft x-ray measurements and machine learning. *J. Space Weather Space Clim.* **11**, 59. DOI.

Andrews, M.D.: 2003, A search for cmes associated with big flares. *Solar Phys.* **218**(1), 261. DOI.

Barnes, G., Leka, K.D., Schumer, E.A., Della-Rose, D.J.: 2007, Probabilistic forecasting of solar flares from vector magnetogram data. *Space Weather* 5(9). DOI.

Bobra, M.G., Ilonidis, S.: 2016, *Astrophys. J.* **821**(2), 127. DOI.

Bobra, M.G., Sun, X., Hoeksema, J.T., Turmon, M., Liu, Y., Hayashi, K., Barnes, G., Leka, K.D.: 2014, The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: sharps – space-weather hmi active region patches. *Solar Phys.* **289**(9), 3549. DOI.

Bobra, M.G., Wright, P.J., Sun, X., Turmon, M.J.: 2021, Smarps and sharps: two solar cycles of active region data. *Astrophys. J.* **256**(2), 26. DOI.

Breiman, L.: 2001, Random forests. *Mach. Learn.* **45**, 5.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: 2002, Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321.

Chen, T., Guestrin, C.: 2016, Xgboost: a scalable tree boosting system. In: *Proc. 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785.

Chen, Y., Manchester, W.B., Hero, A.O., Toth, G., DuFumier, B., Zhou, T., Wang, X., Zhu, H., Sun, Z., Gombosi, T.I.: 2019, Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters. *Space Weather* **17**(10), 1404.

Cortes, C., Vapnik, V.: 1995, Support-vector networks. *Mach. Learn.* **20**(3), 273. DOI.

Cranmer, S.R.: 2009, Coronal holes. *Living Rev. Solar Phys.* **6**, 1.

Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J.A., Benvenuto, F., Bloomfield, D.S., Georgoulis, M.K.: 2018, Forecasting solar flares using magnetogram-based predictors and machine learning. *Solar Phys.* **293**(2), 28.

Freund, Y.: 1995, Boosting a weak learning algorithm by majority. *Inf. Comput.* **121**(2), 256.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: 2014, Generative adversarial nets. *Adv. Neur. In.* 27.

Gosling, J.T.: 1993, The solar flare myth. *J. Geophys. Res.* **98**(A11), 18937. DOI.

Gosling, J., McComas, D., Phillips, J., Bame, S.: 1991, Geomagnetic activity associated with Earth passage of interplanetary shock disturbances and coronal mass ejections. *J. Geophys. Res.* **96**(A5), 7831.

Harrison, R.A.: 1995, The nature of solar flares associated with coronal mass ejection. *Astron. Astrophys.* **304**, 585. ADS.

Hunter, J.D.: 2007, Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* **9**(3), 90.

Inceoglu, F., Jeppesen, J.H., Kongstad, P., Marcano, N.J.H., Jacobsen, R.H., Karoff, C.: 2018, Using machine learning methods to forecast if solar flares will be associated with cmes and seps. *Astrophys. J.* **861**(2), 128.

Kasapis, S., Zhao, L., Chen, Y., Wang, X., Bobra, M., Gombosi, T.: 2022, Interpretable machine learning to forecast sep events for solar cycle 23. *Space Weather* **20**(2), e2021SW002842. DOI.

Leka, K.D., Barnes, G.: 2007, Photospheric magnetic field properties of flaring versus flare-quiet active regions. IV. A statistically significant sample. *Astrophys. J.* **656**(2), 1173. DOI.

Lin, J., Forbes, T.G.: 2000, Effects of reconnection on the coronal mass ejection process. *J. Geophys. Res.* **105**(A2), 2375. DOI.

Liu, C., Deng, N., Wang, J.T., Wang, H.: 2017, Predicting solar flares using sdo/hmi vector magnetic data products and the random forest algorithm. *Astrophys. J.* **843**(2), 104.

Liu, H., Liu, C., Wang, J.T., Wang, H.: 2019, Predicting solar flares using a long short-term memory network. *Astrophys. J.* **877**(2), 121.

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., Shin, G.: 2018, Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms. *Astrophys. J.* **869**(2), 91.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: 2011, Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825.

Pesnell, W.D., Thompson, B.J., Chamberlin, P.C.: 2012, The solar dynamics observatory (SDO). *Solar Phys.* **275**(1–2), 3. DOI. ADS.

Priest, E., Forbes, T.: 2002, The magnetic nature of solar flares. *Astron. Astrophys. Rev.* **10**(4), 313.

Qahwaji, R., Colak, T., Al-Omari, M., Ipson, S.: 2008, Automated prediction of cmes using machine learning of cme – flare associations. *Solar Phys.* **248**(2), 471. DOI.

Schou, J., Scherrer, P.H., Bush, R.I., Wachter, R., Couvidat, S., Rabello-Soares, M.C., Bogart, R.S., Hoeksema, J.T., Liu, Y., Duvall, T.L., Akin, D.J., Allard, B.A., Miles, J.W., Rairden, R., Shine, R.A., Tarbell, T.D., Title, A.M., Wolfson, C.J., Elmore, D.F., Norton, A.A., Tomczyk, S.: 2012, Design and ground calibration of the helioseismic and magnetic imager (HMI) instrument on the solar dynamics observatory (SDO). *Solar Phys.* **275**(1–2), 229. DOI. ADS.

Schrijver, C.J.: 2009, Driving major solar flares and eruptions: a review. *Adv. Space Res.* **43**(5), 739. DOI. https://www.sciencedirect.com/science/article/pii/S0273117708005942.

Shibata, K., Magara, T.: 2011, Solar flares: magnetohydrodynamic processes. *Living Rev. Solar Phys.* **8**, 1.

Sun, P., Dai, W., Ding, W., Feng, S., Cui, Y., Liang, B., Dong, Z., Yang, Y.: 2022, Solar flare forecast using 3d convolutional neural networks. *Astrophys. J.* **941**(1), 1. DOI.

Tharwat, A., Gaber, T., Ibrahim, A., Hassanien, A.E.: 2017, Linear discriminant analysis: a detailed tutorial. *AI Commun.* **30**(2), 169. DOI.

Torres, J., Zhao, L., Chan, P.K., Zhang, M.: 2022, A machine learning approach to predicting sep events using properties of coronal mass ejections. *Space Weather* **20**(7), e2021SW002797. DOI.

Wang, J., Liu, S., Ao, X., Zhang, Y., Wang, T., Liu, Y.: 2019, Parameters derived from the sdo/hmi vector magnetic field data: potential to improve machine-learning-based solar flare prediction models. *Astrophys. J.* **884**(2), 175.

Wang, X., Chen, Y., Toth, G., Manchester, W.B., Gombosi, T.I., Hero, A.O., Jiao, Z., Sun, H., Jin, M., Liu, Y.: 2020, Predicting solar flares with machine learning: investigating solar cycle dependence. *Astrophys. J.* **895**(1), 3. DOI. ADS.

Webb, D.F., Howard, T.A.: 2012, Coronal mass ejections: observations. *Living Rev. Solar Phys.* **9**(1), 3. DOI.

Woodcock, F.: 1976, The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Weather Rev.* **104**(10), 1209.

Yan, X.-L., Qu, Z.-Q., Kong, D.-F.: 2011, Relationship between eruptions of active-region filaments and associated flares and coronal mass ejections. *Mon. Not. Roy. Astron. Soc.* **414**(4), 2803. DOI.

Yashiro, S., Gopalswamy, N.: 2009, Statistical relationship between solar flares and coronal mass ejections. In: Gopalswamy, N., Webb, D.F. (eds.) *Universal Heliophysical Processes* **257**, 233. DOI. ADS.

Yi, K., Moon, Y.-J., Lim, D., Park, E., Lee, H.: 2021, Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters. *Astrophys. J.* **910**(1), 8.

Zhang, H., Li, Q., Yang, Y., Jing, J., Wang, J.T., Wang, H., Shang, Z.: 2022a, Solar flare index prediction using sdo/hmi vector magnetic data products with statistical and machine-learning methods. *Astrophys. J.* **263**(2), 28.

Zhang, H., Li, Q., Yang, Y., Jing, J., Wang, J.T.L., Wang, H., Shang, Z.: 2022b, Solar flare index prediction using sdo/hmi vector magnetic data products with statistical and machine-learning methods. *Astrophys. J.* **263**(2), 28. DOI.

Zheng, Y., Li, X., Si, Y., Qin, W., Tian, H.: 2021, Hybrid deep convolutional neural network with one-versus-one approach for solar flare prediction. *Mon. Not. Roy. Astron. Soc.* **507**(3), 3519.

Zirker, J.B.: 1977, Coronal holes and high-speed wind streams. *Rev. Geophys.* **15**(3), 257.