Interpretable Machine Learning to Forecast SEP Events for Solar Cycle 23

Spiridon Kasapis[1], Lulu Zhao[2], Yang Chen[3], Xiantong Wang[2], Monica Bobra[4], Tamas Gombosi[2]

[1]Department of Naval Architecture and Marine Engineering, University of Michigan
[2]Department of Climate and Space Sciences and Engineering, University of Michigan
[3]Department of Statistics, University of Michigan
[4]Hansen Experimental Physics Laboratory, Stanford University

Key Points:

- In an experimental setting, SMARP data can correctly predict whether a solar flare will lead to a solar energetic particle (SEP) event 72% of the times.
- Flare peak intensity is the strongest SEP predictor and can be coupled with SMARP data to achieve accuracy $\leq 0.92 \pm 0.07$.
- The SMARP dataset provides a leading time of 55.3±28.6 minutes for forecasting the SEP events.

Corresponding author: Spiridon Kasapis, skasapis@umich.edu

Abstract

We use machine learning methods to predict whether an active region (AR) which produces flares will lead to a solar energetic particle (SEP) event using Space-Weather Michelson Doppler Imager (MDI) Active Region Patches (SMARPs). This new data product is derived from maps of the solar surface magnetic field taken by the Michelson Doppler Imager (MDI) aboard the Solar and Heliospheric Observatory (SOHO). We survey the SMARP active regions associated with flares that appear on the solar disk between June 5, 1996 and August 14, 2010, label those that produced SEPs as positive and the rest as negative. The AR SMARP features that correspond to each flare are used to train two different types of machine learning methods, the support vector machines (SVMs) and the regression models. The results show that the SMARP data can predict whether a flare will lead to an SEP with accuracy (ACC) $\leq 0.72 \pm 0.12$ while allowing for a competitive leading time of 55.3 ± 28.6 minutes for forecasting the SEP events.

Plain Language Summary

We train machine learning algorithms to predict whether sunspots on the solar disk will produce solar phenomena such as Solar Energetic Particle (SEP) events. We use a newly published piece of solar data (SMARP) captured by the Space-Weather Michelson Doppler Imager (MDI), an instrument aboard the Solar and Heliospheric Observatory (SOHO) which is a spacecraft launched on 1995 to study the Sun. The results show that the SMARP data can correctly predict 72% of the times whether a sunspot will lead to an SEP while allowing for a potentially competitive forecast window.

1 Introduction

Large solar eruptions can potentially harm modern civilization in several different ways. Events such as large solar flares and coronal mass ejections (CMEs) that lead to solar particle acceleration, can adversely affect the near-earth environment, degrade high frequency (HF) radio communications, incapacitate satellites, expose airline passengers to elevated radiation levels and even endanger life in outer space. Therefore, predicting and monitoring such events is an important task for the community.

Solar Energetic Particles are rare events that involve protons, electrons and heavy ions accelerated to high energies (up to tens of GeV while the fastest ones can accelerate to speeds of up to 80% of the speed of light) by two solar processes (Reames, 2013), the energization at a solar flare site or the shock waves associated with Coronal Mass Ejections (CMEs). Solar charged particles are accelerated in flares or CME shock waves (Wild et al., 1963) and travel preferentially along the interplanetary magnetic field to their detection point in space (McCracken & Ness, 1966).

The study of solar energetic particle (SEP) events is a relatively recent science as the identification of the first event took place on 28 February, 1942 (Forbush, 1946). Observations of solar proton events (subset of solar energetic particle events) were made using ground-based instruments that detected ionization, neutrons, or radio disturbances caused by them. The largest solar proton event recorded using these modern techniques (particles exceeded 15 GeV at the top of the atmosphere) was on the 23rd of February, 1956. In the mid-1960s spacecraft were deployed that began directly measuring solar proton events. This was also the time when the first flare was associated with an SEP event (Shea & Smart, 1995).

During the so-called Halloween storms in late October 2003, SEP events caused 47 satellites to report malfunctions, more than 10 satellites to go out of action for days, the Mars Odyssey spacecraft went into deep safe mode (Lopez et al., 2004), a Japanese satellite costing 640m USD was completely lost, the US FAA issued their first-ever high ra-

diation dosage alert for high-altitude aircraft, and astronauts in the ISS had to seek safety into their heavily shielded service module (Webb & Allen, 2004; Horne et al., 2013).

One of the sources of solar activity phenomena that cause SEPs are the magnetically strong regions on the solar sphere that we refer to as active regions (van Driel-Gesztelyi & Green, 2015). The most flare productive active regions (ARs) are the ones that undergo large changes in sunspot area and show magnetic flux imbalance (Choudhary et al., 2013). Large active regions are also generally strong, produce a number of flares, evolve rapidly and their lifetime spans from days to months (Choudhary et al., 2013). Using instruments carried onboard satellites such as the Michelson Doppler Imager (MDI) on the Solar and Heliospheric Observatory (SOHO) or the Helioseismic and Magnetic Imager (HMI) on the Solar Dynamics Observatory (SDO), we are able to retrieve components of the magnetic field at the solar surface, allowing us to calculate physical characteristics of the ARs (Scherrer et al., 1995; Schou et al., 2012).

Solar particle prediction studies mainly use the flare and nearEarth space environment data to forecast SEP events given the knowledge that large SEPs are almost always accompanied by a flare (Schrijver et al., 2012). Laurenza et al. (2009) used data such as flare longitude, time-integrated soft X-ray intensity, and time-integrated intensity of type III radio emission at around 1 MHz to provide short-term warnings for SEP events. Similarly, Núñez (2011) used the soft Xray, differential and integral proton fluxes data to forecast the SEP events of Solar Cycle 23 recorded on the NOAA/SWPC list. Although both flare and CME data are found to be useful inputs to predictive models, García-Rigo et al. (2016) deemed it sufficient to only use flare properties as they noticed that the CME information offers insignificant increase in SEP prediction accuracy, because of the difficulty to obtain real-time information on the true radial CME speeds due to line-of-sight issues.

Recently, machine learning (ML) methods like neural networks (in the multi-layer perceptron implementation), random forests, decision trees, extremely randomized trees and other, have been used in predicting SEP events. The preliminary results obtained by Bain et al. (2018) show that machine learning classification techniques such as the logistic regression (LR), decision trees (DTs) and support vector machine (SVM) algorithms give an improved forecasting skill over the current SWPC Proton Prediction Model (Balch, 2008) based on physical parameters associated with solar flares and coronal mass ejections. An even more comprehensive study that assesses the predictability of Solar Energetic Particles using ML techniques was recently published by Lavasa et al. (2021). In their work they conclude that random forests (RF) could be the prediction technique of choice for an optimal sample comprised by both flares and CMEs while proving that the most important features are the CME speed, width and flare soft X-ray (SXR) fluence. Lastly, Sadykov et al. (2021) recently indicated the possibility of developing robust all-clear SPE forecasts by employing machine learning methods. Their approach indicates that for AR-based predictions, it is necessary to take into account western limb and far-side ARs, characteristics of the preceding proton flux represent the most valuable input for prediction, daily median characteristics of ARs and the counts of type II, III, and IV radio bursts may be excluded from the forecast and that ML-based forecasts outperform SWPC NOAA forecasts.

Different studies have used a variety of sources to obtain the data necessary for solar particle event prediction. Richardson et al. (2018) predict the SEP events peak proton intensity at one energy interval (1424 MeV) using the CME data (speed, width, direction and type II and type III radio emissions associated with the CME) in the Space Weather Database of Notications, Knowledge, Information (DONKI). Papaioannou et al. (2016) have presented a catalogue which includes solar variables (such as logarithm of the solar flare magnitude (log SXRs), solar flare longitude, duration, and rise time) for 314 SEP events obtained from the Energetic Particle Sensor (EPS) aboard the Geostationary Operational Environmental Satellites (GOES; Rodriguez et al., 2014) and CME

data (width/size and velocity) obtained by the Large Angle and Spectrometric Coronagraph (LASCO; Brueckner et al., 1995) carried onboard the SOHO spacecraft. Using this information, Papaioannou et al. (2018) classify the solar energetic particle (SEP) event radiation impact with respect to the characteristics of their parent solar events while attempting to infer the possible prediction of SEP events.

Similarly, Anastasiadis et al. (2017) provide full-disk Helioseismic and Magnetic Imager (HMI) magnetograms to their novel integrated prediction system which nowcasts SEP events. The HMI instrument aboard the Solar Dynamics Laboratory (SDO) measures the solar surface magnetic field from which the Space-Weather HMI Active Region Patches (SHARPs) are derived. SHARPs have been used to identify flares or SEPs in Chen et al. (2019) and Inceoglu et al. (2018) respectively.

A new data product recently published by Bobra et al. (2021) called Space-Weather MDI Active Region Patches (SMARPs) will be used in this work to predict SEPs. SMARPs are derived from the solar surface magnetic field taken by the Michelson Doppler Imager (MDI) on the SOHO spacecraft and provide a continuous and seamless set of keywords that describe every active region observed during Solar Cycle 23. The big difference between the HMI (Schou et al., 2012) and the MDI (Scherrer et al., 1995) is that the first measures the vector magnetic field at the solar surface whereas the later only measures the line-of-sight component of the solar magnetic field. The main aim of this study is to evaluate the predictive power of MDI Active Region Patches (SMARPs) on SEP events as it is desirable for the space weather community to explore new datasets that, when used on machine learning algorithms in the future, will be able to predict when solar proton events will occur, along with how energetic and how intense they will be.

2 Database

In this work, we will evaluate the predictive power of the MDI solar magnetogram on SEP events. In particular, we focus on whether an active region which is associated with a solar flare will lead to an SEP event. Compared to other works that predict whether an SEP is likely to occur in a defined future time window, our models simply forecast whether flares have a resulting particle increase. To achieve this, five different predictors obtained from the SMARP dataset (SMARP Predictors) are used, while two more predictors from the NOAA solar X-ray flare dataset (Flare Predictors) are used for comparison. While we are specifically interested in the responses of the ML models when only SMARP Predictors are used, the ability to forecast SEPs by using flare data will serve as a baseline capability.

2.1 SMARP Predictors

The magnetogram is measured by the Michelson Doppler Imager (MDI Scherrer et al., 1995) onboard SOHO between June 5, 1996 and August 14, 2010. Based on the magnetogram, Bobra et al. (2021) derived a new database called Space-Weather MDI Active Region Patches (SMARPs), which contains characteristics of the active regions on the solar surface. A Tracked Active Region Patch (TARP) Number is assigned to each active region as its identification number and a NOAA active region number, if available, is assigned to each active region patch. Three physical keywords, total unsigned flux (USFLUXL), mean gradient of the vertical field (MEANGBL), and the logarithm of the total unsigned flux near polarity inversion line (RVALUE) are calculated using the pixels in the active region and stored in the SMARP header file. In addition, the SMARP header file also contains four spacial features specifying the location of the corresponding AR on the solar surface: the minimum and maximum latitude (LATDMIN, LATDMAX) and the minimum and maximum longitude (LONDTMIN, LONDTMAX). The SMARP data is available on the Joint Science Operations Center database (Mumford et al., 2015; Barnes et al., 2020).

Besides the three physical keywords stored in the SMARP header file, we calculate the angular distance between the AR and the magnetic foot-point of the earth. The longitude and latitude location of the active region on the sun is approximated by the geometric center of the active region using the latitude and longitude keywords. The magnetic foot-point of the earth on the sun is assumed to be at W45°. Note that the magnetic foot-point varies from event to event. One way of characterizing this variability is to calculate the magnetic foot-point location using the solar wind speed measured at 1 AU assuming an ideal Parker spiral up to the solar source surface and reconstruct the coronal magnetic fields using potential field source surface model. However, interplanetary magnetic field can also be disturbed by corotating interaction regions (CIRs), interplanetary coronal mass ejections (ICMEs) and other solar transient events, especially in solar maximum. In this work, for simplicity, we use W45° as an approximation. We also calculate the size of the active region by multiplying the difference of longitude by the difference of latitude.

2.2 Construction of SEP Event List

The SEP event list we use in this work is documented in the NOAA Space Environment Service Center website. This study is based strictly on the Solar Energetic Particle categorization that NOAA SWPC follows, therefore the catalogue lists the SEP events observed by GOES that are accelerated to energies greater than 10 MeV, noting only the times when the proton flux exceeds this threshold. This means that NOAA sometimes groups more than one SEP particle injection together as one event, which is an intrinsic problem of the dataset. Really strong ARs are also known to produce multiple SEP events, therefore there are a number of homologous events coming from the same AR. For each SEP event, a solar flare and the corresponding NOAA active region number is assigned if exists. The solar flare list is obtained from the NOAA Solar Flare Data website. For each solar flare, the list contains the start, peak and end times, the peak intensity of the flare, the active region location and the corresponding NOAA active region number.

We match the solar flare list with the SMARP database using the AR numbers. If a flare does not have a registered AR number, matching based on their occurrence time and spatial coordinates is performed. We also discard those solar flares whose flare class, coordinates or AR numbers are undefined or missing. Out of the $\sim$ 25,000 flares (A, B, C, M, X) recorded during the 14 year span between 1996 and 2010, only 6,510 flares have sufficient information and therefore are matched with SMARP files.

During this 14 year span, 93 SEP events are detected by the GOES spacecraft. Missing information about the SEP's associated flare or AR such as the Location and its Importance (Xray/Opt), leave only 70 SEPs with information adequate to label the 6,510 flares. We assign a label to each flare: Positive if it led to an SEP and Negative if it did not. An additional 5 SEP-flare couples were discarded due to missing physical feature data about their corresponding SMARP Active Region. Therefore, the dataset used for training has a Positive and a Negative component comprised of 65 and 6,510 flares respectively, making it vastly unbalanced.

The SMARP header files contains rows with the physical and spatial features of each active region at a 96-minute cadence throughout its entire lifetime, starting two days before it emerges or rotates onto the solar disk until two days after it submerges or disappears from view behind the limb (Bobra et al., 2021). We select the SMARP header file row at the time right before the flare peak time. In Figure 1, a histogram of the time difference between the selected SMARP file row and the flare peak time is plotted with the left panel corresponding to the positive dataset and right panel corresponding to the negative dataset. The importance of the SMARP time characterization varies between different SMARP active regions. Picking earlier SMARP points can potentially yield slightly

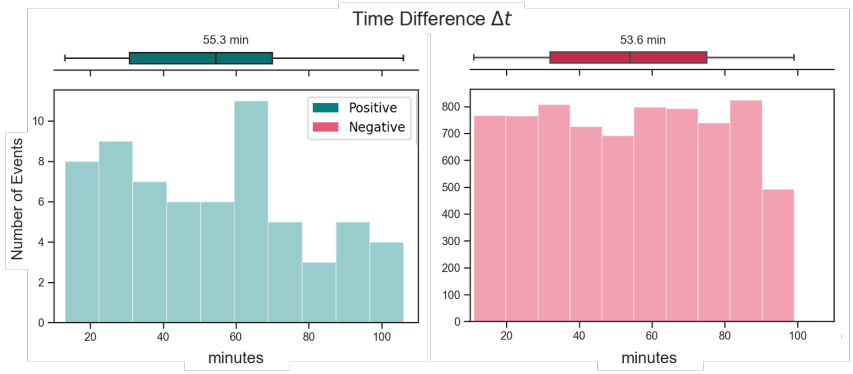

Figure 1. Histograms for the time difference between the flare peak time and the selected SMARP data (selected row in the TARP file). The distributions range between 10 and 100 minutes. The mean time differences shown in the error bars above the graphs are 55.3 and 53.6 minutes with a standard deviation of 28.6 and 24.8 minutes for the Positive (green) and Negative (red) datasets respectively.

different results than the ones presented below but the scope of the paper is to give an evaluation of the SMARP dataset rather than examine the optimal prediction window. The distributions range between 10 and 100 minutes. The mean time difference is 55.3 and 53.6 minutes with a standard deviation of 28.6 and 24.8 minutes for the Positive and Negative dataset respectively. The time difference Δt should not be confused with forecast windows or a lead times presented in similar works as the non-operational nature of our study does not allow for foreknowledge of whether a flare will occur.

2.3 Flare Predictors

We will evaluate the prediction power of SMARP dataset on SEP events by comparing the prediction results with those obtained by only using the flare information, i.e. flare peak intensity and flare location. We use the solar long wavelength X-ray flare data that NOAA's Geostationary Operational Environmental Satellites (GOES) continuously provides since 1975. Similarly to the SMARP Predictors, we calculate the flare angular distance from the earth's magnetic foot-point location, W45°, on the sun.

3 Preliminary Data Analysis

We conduct preliminary analysis/assessment of different predictors, i.e. the SMARP and the Flare Predictors, via comparing the histogram of each predictor for the positive with that of negative samples. Figure 2 shows the density histograms of each predictor from the SMARP dataset on the top and from the GOES flare information on the bottom. The positive data is shown in green and negative data in red.

As shown in Figure 2, the flare peak intensity is a powerful discriminator between the positive and negative dataset. The flare peak intensity has been used as a feature to predict the occurrence and properties (peak proton intensity, event duration, and etc.) of SEP events (Laurenza et al., 2009; Balch, 2008). The four intervals present in the intensity graph ($[-7,-6]$, $[-6,-5]$, $[-5,-4]$ and $[-4,+\infty]$) represent the four different GOES X-ray classes B, C, M and X respectively. The predictive power difference between the flare peak intensity and the SMARP Predictors on the top of Figure 2 has a big impact when comparing the SEP prediction capability with and without SMARP data. Flare intensity is the predictor that has the least overlap between positive and negative. In the

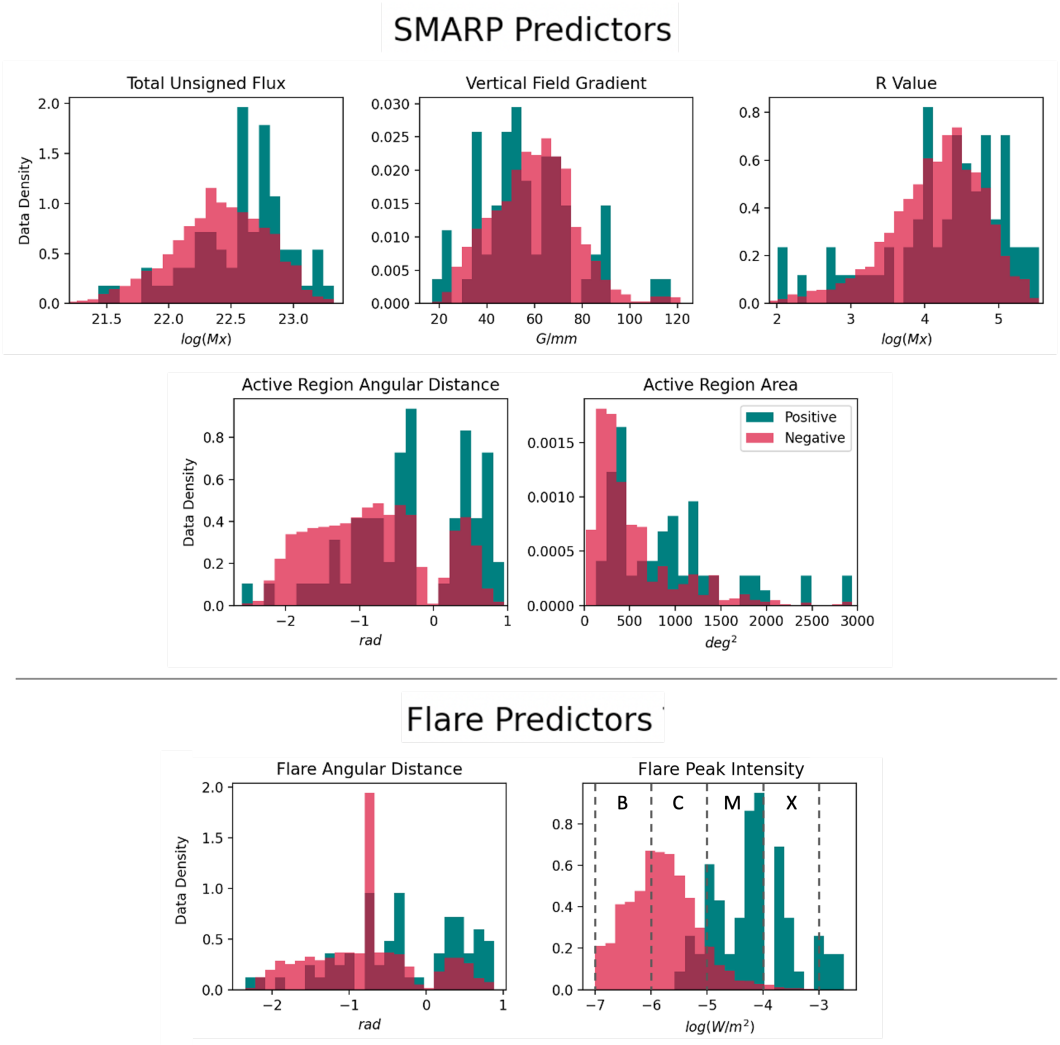

Figure 2. The probability density values are given in the histograms for the SMARP and Flare Predictors such that the area under each histogram integrates to 1. Both the SEP (green, positive samples) and flare-only (red, negative samples) data are split into 25 bins.

positive parameter spaces where there is no overlap, the associated SEP events are very intense and contain higher energy protons. Although larger flare intensity leads to higher proton fluxes, this is not necessarily the rule for all events as the problem is multidimensional (i.e. the correlation coefficient between the two physical quantities is not 1).

Moderate distinction between the events that led to an SEP and those that did not can be identified in the predictors acquired using the SMARP active region coordinates (Active Region Angular Distance and Area). Large active regions increase the likelihood of an SEP event occurrence. The total unsigned flux is related to the X-ray, the EUV emissions from the sun and the particle acceleration, which reflects the energy stored in an active region (Gurman et al., 1974; van Driel-Gesztelyi et al., 2003; Ugarte-Urra et al., 2015), therefore the SEP events are connected to higher flux values. The flux and intensity distributions show similar trends but with the former having less predictive power. The Vertical Field Gradient distribution of the Positive dataset aligns well with that of

the Negative dataset, making it the least powerful predictor along with the R Value which shows the same trend.

The predictive power differences observed in Figure 2 are consistent with a t-test (Kim, 2015) performed to all variables. For the most discriminating predictors, such as the Flare Peak Intensity and the AR Area, the Statistic values are high (21.3 and 6.35 respectively) whereas for the weakest predictors (R Value and Vertical Field Gradient) the Statistic values do not exceed the 1.58 value.

4 Machine Learning Methods

To investigate whether the SOHO (SMARP AR data) or the GOES (flare eruption information) dataset can predict better the response variable of the two classes defined above, we use two popular groups of machine learning algorithms provided by the scikit-learn software package v0.24.2 for Python: different variations of the Support Vector Machine (SVM; Cortes & Vapnik, 1995) and two Regression Models.

4.1 Support Vector Machine

SVMs were initially designed and have been used to solve binary classification problems (Shao et al., 2014). In the most general case, the SVM is fitted to the data using a set of vector-target pairs (x_i, y_i) where $i = 1, 2, .., n$. The target for positive and negative observations respectively is $y_i \in \{1, 0\}$ and the corresponding physical characteristics feature vector is $x_i = (f_{i1}, f_{i2}, .., f_{ip})$. For all tests performed, our training data length is $n = 116$ and the maximum feature vector length is $p = 7$, where all calculated predictors are used. Each different SVM method maps the input feature vector x_i to a higher dimension space using an unknown function ϕ dependent on the user-defined kernel K. Given a regularization parameter $C > 0$ it solves an optimization problem to obtain the SVM trained weight vector w (Hsu & Lin, 2002; Inceoglu et al., 2018). The regularization parameter C controls the scaling of the SVM loss function and compensates for the change in the number of samples between the main problem and the smaller problems within the folds of the cross-validation. During testing, prediction is done by multiplying the trained vector w to the projected input feature vector $\phi(x_i)$ with an addition of a bias term. A more detailed study on how to solve the SVM optimization equations is out of the scope of this research and can be found elsewhere (Cortes & Vapnik, 1995; Vapnik, 1998).

The kernel function K is defined as the inner product of data pairs that correspond to different observations i and j, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. In this study we train four different variations of the SVM (Amari & Wu, 1999). One uses the Linear kernel $K = \langle x_i, x_j \rangle$, two use Polynomial kernels $K = \gamma \langle x_i, x_j \rangle^d$ where $d \in \{2, 3\}$ (second and third degree) and the last one uses the Gaussian Radial Basis Function (RBF) kernel $K = exp(-\gamma \|x_i = x_j\|^2)$ which has been used in similar studies (Inceoglu et al., 2018). To mitigate the risk of overfitting that is anticipated with the RBF kernel, the model parameters are being chosen using the results that the ~~validation~~ testing dataset yields. Using this cross-validation method we get similar results for the RBF kernel as we do for the rest models, which proves that excessive overfitting is being avoided. The weighting factor γ is user-defined and controls the influence a single training example has on the classification task. The different kernels help the prediction model deal with complex datasets such as the physical features of solar active regions by transforming the input into any desired form.

4.2 Linear Models

The observed physical properties of a SMARP AR can be also processed for the purpose of prediction by linear models: regression methods in which the target value is

expected to be a linear combination of the input features. Assuming a model function $f(x) = w^T x + b$ where w is a set of coefficients acquired during fitting, every feature's (x_i) predicted target y_i is 1 if $f(x_i) \geq 0$ and 0 if otherwise. In this case study linear models such as Ridge and Logistic regression are being used.

The ridge regression is one of the simplest machine learning algorithms and works well for small datasets while being computationally inexpensive. To fit the coefficients w to the training data, the ridge regression minimizes an ordinary Least Squares loss function with an additional term that penalizes the size of the coefficients, as given in (1).

$$J_{Ridge} = \|w^\top x - y\|_2^2 + \alpha \|w\|_2^2 \tag{1}$$

Between different training runs we vary the complexity parameter α in order to control the amount of shrinkage and find the value that produces the most robust predictions. More specifically, α corresponds to $1/(2C)$ in the Logistic Regression or SVM and serves as a regularization parameter which improves the conditioning of the problem and reduces the variance of the estimates. Larger values specify stronger regularization. We do not adopt a cross-validation procedure for selecting the tuning parameter α due to considerations of sample sparsity and because the randomized picking process of the training data leads to non-significant selection bias. Although it is a model often adopted when the response y takes real numbers, we chose ridge regression because it reduces overfitting, guarantees that we can find a solution and offers a different approach for binary classification compared to other competing models.

The dichotomous nature of Logistic Regression makes it a great candidate for the binary SEP prediction task. We use the default Logistic Regression module provided by the Scikit-Learn library in Python (Pedregosa et al., 2011) which includes the l_2 regularization as a penalty and the Limited-memory BroydenFletcherGoldfarbShanno (L-BFGS) optimization algorithm (Saputro & Widyaningsih, 2017) as a solver. The L-BFGS solver fits our application as it is robust and recommended for small dataset prediction tasks. To calculate the optimal w coefficients, Logistic Regression minimizes the cost function J for w and c.

$$J_{LR} = \frac{1}{2} w^\top w + C \sum_{i=1}^{n} \log(e^{-y_i (x_i^\top w + c)} + 1) \tag{2}$$

The constant C controls the regularization strength of the model. Although normalization is applied to the flare data before the fitting process, the C constant is also varied throughout different training runs in order to find the value that produces the most numerically stable prediction.

4.3 Training and Tuning the ML Models

The scarcity of the SEP events along with the mission duration of the MDI/SOHO limits the size of the Positive dataset and leads to difficulties in separating the data into training and testing subsets in a reasonable way. To overcome this problem, every model is trained on 90% of the Positive events (58) and an equal number of Negative events. The training of each algorithm is followed by a similarly balanced testing on the remaining 10% Positive (7) and an equal number of Negative events. This balanced training and testing procedure is repeated k number of times to provide uncertainty assessment of the random selection of events. In our work, k is chosen to be equal to 100. Each time, a different batch of Negative events is randomly selected from the pool of 6,510 flare eruption events that did not lead to an SEP. This means that the Negative events are selected without replacement (in every one of the k different runs, all negative train and test samples are different) while the Positive class is selected with replacement due to the low

number of available events. Similarly, in every run a different split between training and testing occurs for the Positive dataset. It has to be highlighted that as the described methodology dictates, there is at no point any overlap between training and testing, neither in the positive nor in the negative dataset. The training dataset is independent from the testing therefore throughout the paper we only present the testing results. Furthermore, the procedure that we describe above for assessing the uncertainty of the testing results follows the idea of bootstrap (Efron, 1979; Efron & Tibshirani, 1994), which uses random Monte Carlo samples to assign measures of accuracy, i.e. variance in our case. Across the many (k=100) resampling of the training and testing sets, it is with probability almost 1 that we cover the best and worst cases in terms of testing accuracy and other metrics. Therefore, the uncertainty assessment that we give in this paper is crucial in understanding and interpreting our model performance.

It is important to acknowledge that the prediction of such rare events is a very difficult task, therefore our work is focused on using the SMARP dataset to underline the contrast between the flares that lead to SEPs and flares that do not. Using the aforementioned machine learning models we aim to understand the mechanism of SEP formation and explore prediction capabilities rather than presenting an apparatus ready for operation. More specifically, the results presented below aim to evaluate the potential the SMARP dataset has on predicting SEPs while showing which physical parameters are the most important for SEP prediction.

The very important limitations this work faces should also be noted. A very small number of positive data along with the large uncertainties inherited by the SMARP dataset are problems addressed by running the aforementioned Monte Carlo experiments (repeated random sampling) and by quantifying the uncertainties using box plots, statistical spreads and standard deviations as seen in Section 5. Different solar cycles have different numbers of events and properties, therefore another limitation is that we are only able to use data from solar cycle 23. Models that are ready for operational use and are trained on highly unbalanced datasets need an increased weight to the penalty of the loss function for the rare classes of data. In the models presented in this work such weighting is not included, therefore it is expected that the model -if deployed as is- will have an increased number of false alarms, as discussed in Section 5.4.

5 Results

For each one of the SVMs and Linear Models, we follow the same training procedure, aiming to predict whether an AR that produces a flare will lead to an SEP event. The goal is to illustrate how useful the SMARP dataset is for this particular task, we therefore train the ML models using two separate sets of features, one that uses SMARP information and one that uses flare information (see Section 2 for the detailed descriptions of the two sets of features). The number of features vary from 2 to 5 and the machine learning algorithms are tested on a number of different predictor combinations.

The comparison between the different types of predictors and algorithms is done using three metrics that characterize and quantify the predictive power of classifiers: the Accuracy (ACC), the True Skill Statistics (TSS; Hanssen & Kuipers, 1965) and the Heidke Skill Score (HSS; Heidke, 1926). The TSS and the HSS can take any values between -1 (all incorrect) and 1 (all correct) while a value of 0 indicates a random forecast. Similarly, the ACC can take values between 0 and 1 with 0.5 being the score of a random forecast. More information about the metrics, their equations and statistical meaning can be found in the works of Inceoglu et al. (2018) and Florios et al. (2018). For every set of k different runs a cumulative contingency table (Figure 3, 4 and 5) is obtained based on the results from the raw SVM and Linear Model outputs. Each row and column in the contingency table represents the number of instances in an actual class and in a pre-

dicted class respectively. In an ideal case, the main diagonal of a contingency table gets high values while the rest of the matrix gets small values.

The ACC can be artificially high in the rare event where a model always predicts the majority class. In flare and SEP prediction such naive cases are common due to the data imbalance, but in this paper all models are trained on a one-to-one positive-negative ratio, so these rare cases are not a concern. Therefore, this study's basic prediction quality metric is the ACC, with the TSS and HSS being presented too as auxiliary metrics.

5.1 SEP Prediction with SMARP Predictors

The cumulative contingency table in Figure 3 shows that out of 1400 testing instances, Third Degree Polynomial SVM correctly classifies 552 as being Positive and 449 as being Negative when using the SMARP Flux and the AR Distance. This is the cumulative information obtained from 100 different runs, each of which has 14 testing points, therefore out of the 1400 total flares examined, there were 148 false alarms (flares wrongfully being classified as they will lead to an SEP) and 251 missed events (flares that lead to SEPs but the model predicted they do not). The ratio of all correctly classified positive and negative events over the total number of events is the accuracy (ACC). In this case, the mean accuracy suggests that 72% of the times ($\pm$12% for a single run) the Third Degree Polynomial SVM algorithm can predict whether a flare will lead or not to an SEP using the its AR SMARP features. The comparison between the probability of detection and the probability of false detection lead to an average TSS level of 0.47 ± 0.24 for the same 100 runs. Similarly, the HSS measures a fractional forecast improvement over a random forecast of 0.44 ± 0.25.

The results show that the Linear Models can predict whether a flare will be accompanied by SEPs with ACC values $\geq 0.70 \pm 0.12$ for a number of SMARP Predictor combinations (Table 2 in Appendix). The maximum corresponding TSS and HSS values for these SMARP Predictor runs are above the 0.40 levels. Similar to row 3 of Table 3, the Polynomial models that use the AR Distance and Area in Table 2 fail to produce a meaningful decision boundary yielding ACC values $\leq 0.52 \pm 0.04$, TSS values $\leq 0.15 \pm 0.28$ and HSS values $\leq 0.05 \pm 0.09$. Note here that a zero TSS or HSS value means that the method has no skill over the random forecast, therefore the these specific Polynomial examples do not show any predictive power at all.

Although the quality of the results cannot be judged based on the variance, the scores indicate that the better a model's predictive power is, the lower the variance between the different runs is. Thus, the flare peak intensity based models (Table 3, row 6-10) have an ACC standard deviation of ≤ 0.09 while in SMARP examples, where prediction quality is inferior, the ACC standard deviation is ≥ 0.10. This pattern is even more evident when considering the TSS (or the HSS) for which the standard deviation can be as high as 0.28 at the SMARP Predictors exclusive runs in Table 2. Potential reasons about this behavior of variance is the small number of Positive data which allows for low quality runs to not converge at all.

Both SVM and Linear models are affected by user-defined constants such as the α and C in Equations 1 and 2. An embedded grid search -using the testing dataset to avoid overfitting- is employed for each experiment, where we vary each hyper-parameter on a range between 0.05 to 20. The parameter that produced the highest-quality and most consistent results was used for the examples presented in this study.

The negative dataset includes SMARP data only from the active regions that produced at least one flare throughout their lifespan. In a real-world application, the forecaster does not have prior knowledge of whether an AR is going to produce flares or not. In this work we chose to disregard quiet active regions (ARs that did not flare) not only because they are missing the flare predictors we compare SMARPs with, but mainly be-

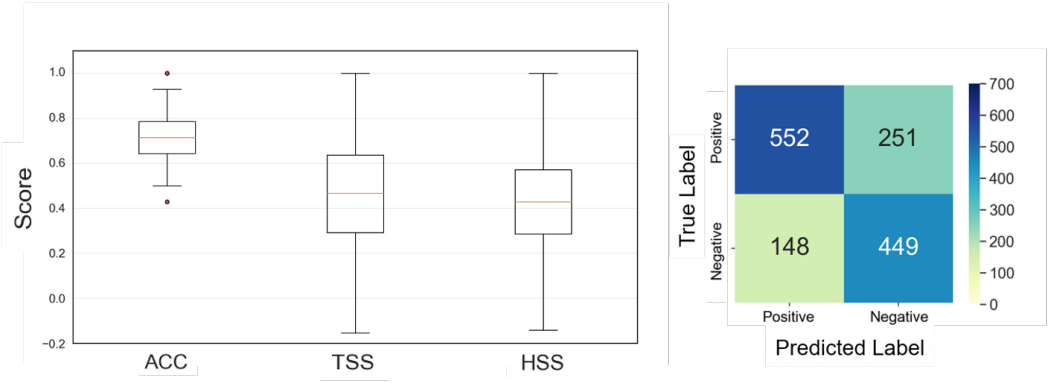

Figure 3. The distribution of k = 100 different ACC, TSS and HSS values are shown in the box plots (left). The values were obtained using the USFLUX & AR Distance on a Third Degree Polynomial SVM and constitute the best SEP prediction the SMARP data can achieve. The box range shows the interquartile range, the red line inside it the median value, the whiskers show the results range and the two red dots show two outlier values. The range of the y-axis is kept the same with Figure 5 for comparison. Adding all the individual TP, TN, FP and FN values respectively we produce a cumulative contingency table for the 100 different runs (right). For the runs where only SMARP predictors are used, the contingency table yields a False Alarm Ratio (FAR) of 0.211 and a Probability of Detection (POD) of 0.687.

cause they are very easily distinguishable compared to the ARs that produced SEPs. This claim is demonstrated in Figure 4, where the results of the binary classification between Positive SMARP events and SMARPs chosen randomly from quiet active regions are presented.

The quiet dataset was created using the 3901 SMARPs that do not produce flares. Out of these 3901 ARs, almost half have missing data (especially RVALUE and USFLUX values that are zero). From the ones that have an adequate amount of data, we randomly select a SMARP data point, creating a quiet dataset of 1529 different predictor vectors.

In the histograms of Figure 4 we observe that the overlap between quiet and positive datasets for the USFLUX, ARAREA and RVALUE predictors is insignificant. The results under the histograms show that the classification task between positive and quiet, when using the USFLUX, the ARAREA and the RVALUE as predictors on an Polynomial SVM model, yield mean values of 0.96 ± 0.05, 0.93 ± 0.09 and 0.91 ± 0.10 for the ACC, TSS and HSS respectively. These numbers very well demonstrate that distinguishing between Positive and Quiet active regions is a trivial problem for our machine learning models, yielding classification accuracy errors of less than 4%.

5.2 SEP Prediction with Flare Predictors

Similar to the results presented before, the prediction quality metrics for the flare-only cases are calculated based on the contingency tables obtained from each different run. The cumulative contingency table in Figure 5 shows that out of 1400 testing instances, Ridge Regression correctly classifies 626 as being Positive and 651 as being Negative when trained on flare peak intensity and distance. This is the cumulative information obtained from $k = 100$ different runs, each of which was tested on 14 data points. The mean accuracy suggests that 91% of the times ($\pm 8\%$ for a single run) the Ridge algorithm can predict whether a flare will lead or not to an SEP using its physical characteristics. The comparison between the probability of detection and the probability of false detection

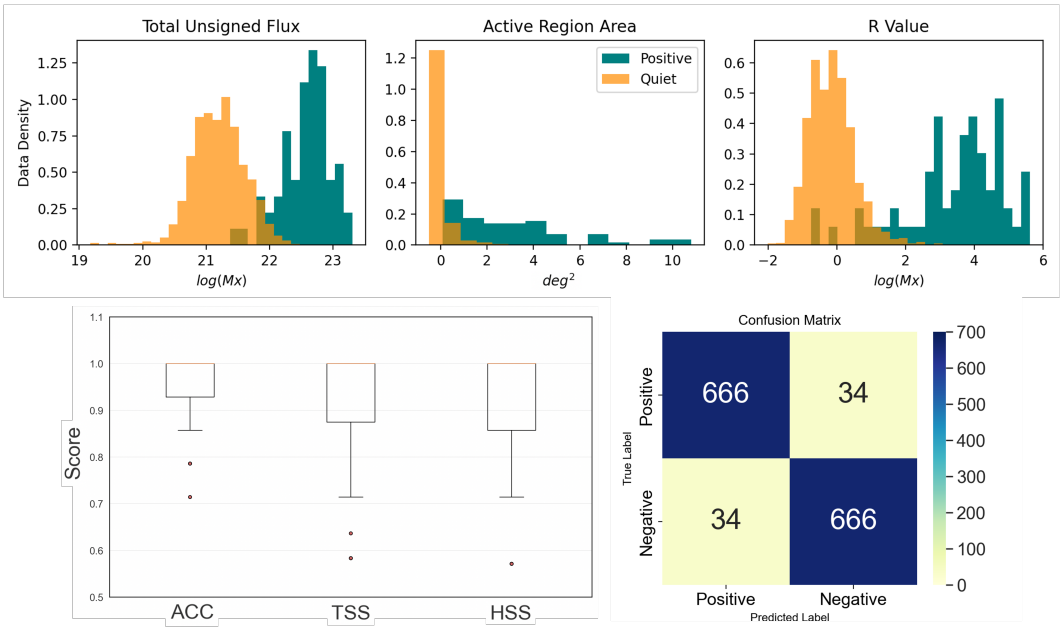

Figure 4. On the top panels of the figure, the probability density values are given in the histograms for three SMARP predictors (USFLUX, ARAREA and RVALUE) such that the area under each histogram integrates to 1. Both the SEP (green, positive ARs) and non-flaring (orange, quiet ARs) data are split into 25 bins. On the bottom part of the figure, the boxplots of the skill scores along with the cumulative contingency table for 100 different runs are presented. We observe that the median values for all three skill scores (ACC, TSS and HSS) are 1, which means no-error classification. For the runs that concern the quiet active regions, the contingency table yields a False Alarm Ratio (FAR) of 0.049 and a Probability of Detection (POD) of 0.951.

lead to an average TSS level of 0.84 ± 0.12 for the 100 runs. Similarly, the HSS measures a fractional forecast improvement over a random forecast of 0.82 ± 0.14.

Using explicitly the Flare Predictors (first row of Table 3) all six models produce similar results. The TSS and HSS show higher standard deviation values (varying from 0.13 to 0.17) compared to the ACC. The predictive power of flare peak intensity is demonstrated when comparing the first two box plots in Figure 6 with the third and fourth one, where predictors (such as USFLUX, RVALUE, ARDIST, MEANGBZ etc.) other than Intensity are being used instead.

The ACC, TSS and HSS values range from 0.88±0.09 to 0.92±0.07 (values marked red and green in Table 3 of the Appendix), 0.78±0.17 to 0.86±0.13 and 0.76±0.18 to 0.84±0.15 respectively for the runs that include Intensity accompanied with a SMARP Predictor (row 6-10). These results show that all models, when using the Flare Peak Intensity, can successfully predict $\leq 92\%$ of the times if a flare will be accompanied with an SEP. When using the SMARP Predictors along with the Flare Distance (row 2-5) instead, the ACC values range from 0.60 ± 0.09 to a maximum of 0.71 ± 0.10, the TSS from 0.36 ± 0.3 to 0.846 ± 0.2 and the HSS from 0.19 ± 0.18 to 0.42 ± 0.2. This proves that when Intensity is not involved in the prediction process, all models yield inferior results, losing at the best case 0.17 ± 0.09 from the accuracy metric. We only test our models on a mix of SMARP and Flare Predictors to verify the prediction power of the Intensity, as in real-life applications the two groups of predictors cannot be used together

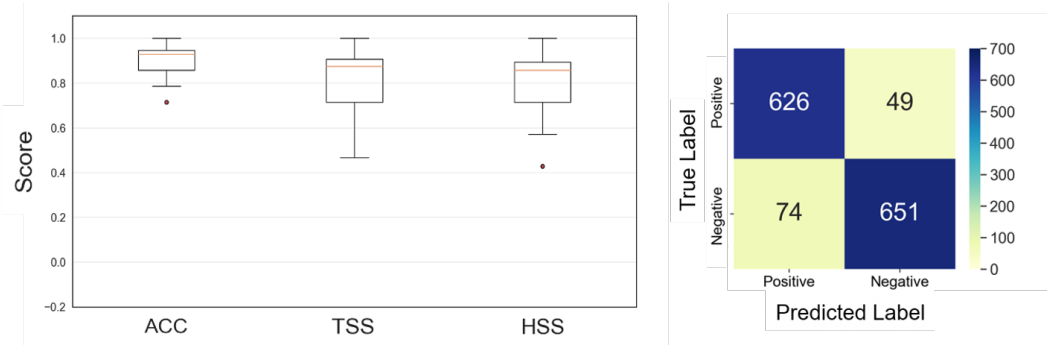

Figure 5. The distribution of k=100 different ACC, TSS and HSS values are shown in the box plots (left). The values were obtained using the Flare Peak Intensity & Flare Distance on a Ridge Regression model and constitute the best SEP prediction the flare data can achieve. The box range shows the interquartile range, the red line inside it the median value, the whiskers show the results range and the two red dots show two outlier values. The range of the y-axis is kept the same with Figure 3 for comparison. Adding all the individual TP, TN, FP and FN values respectively we produce a cumulative contingency table for the 100 different runs (right). For the runs where only flare predictors are used, the contingency table yields a False Alarm Ratio (FAR) of 0.110 and a Probability of Detection (POD) of 0.927.

due to the leading time difference (i.e. SMARP data is being acquired before the flare peak time).

Although each SVM or Linear model performs differently when trained on the same predictors, the variance between the models is of high significance only for some cases where the second and third degree Polynomial SVMs encounter convergence difficulties. The accuracy difference between the best and the worst performing models in Table 3 is always $\leq 0.03 \pm 0.13$, regardless the predictors combination, except for the extreme case of Flare Distance & ARAREA. This consistency gives us confidence to our results while suggesting that our models are not overfitting.

The maximum accuracy achieved on each one of the four main categories of predictor combinations is presented in Figure 6. The resulting ACC, TSS and HSS values show that regardless the machine learning model, the Flare Predictors generally perform better than the SMARP data because of the better predictive power of the flare peak intensity. Although the SMARP data cannot provide SEP forecast of quality similar to the flare peak intensity, it provides us with a larger leading time compared to the Flare Predictors as the flares precede in time the SMARP data points.

5.3 Comparison with Results in Literature

Inceoglu et al. (2018) used data provided by the SHARPs, GOES, and DONKI databases to train SVMs that forecast both CME and SEP events with maximum TSS and HSS of 0.92±0.09 and 0.92±0.08. Anastasiadis et al. (2017) use the SDO/Helioseismic and Magnetic Imager (HMI) full-disk magnetograms and the flare information from the SOHO/MDI database on the prediction tool they call Forecasting Solar Particle Events and Flares (FORSPEF). They achieve Heidke Skill Scores (HSS) of 0.37 ± 0.011 and 0.67 ± 0.007 when using solar flare data and CME data respectively. While we only use GOES data to forecast exclusively SEP events (not CMEs), the best TSS and HSS our SVM implementations achieve are 0.84±0.12 and 0.82±0.14, results that are comparable to both aforementioned studies.

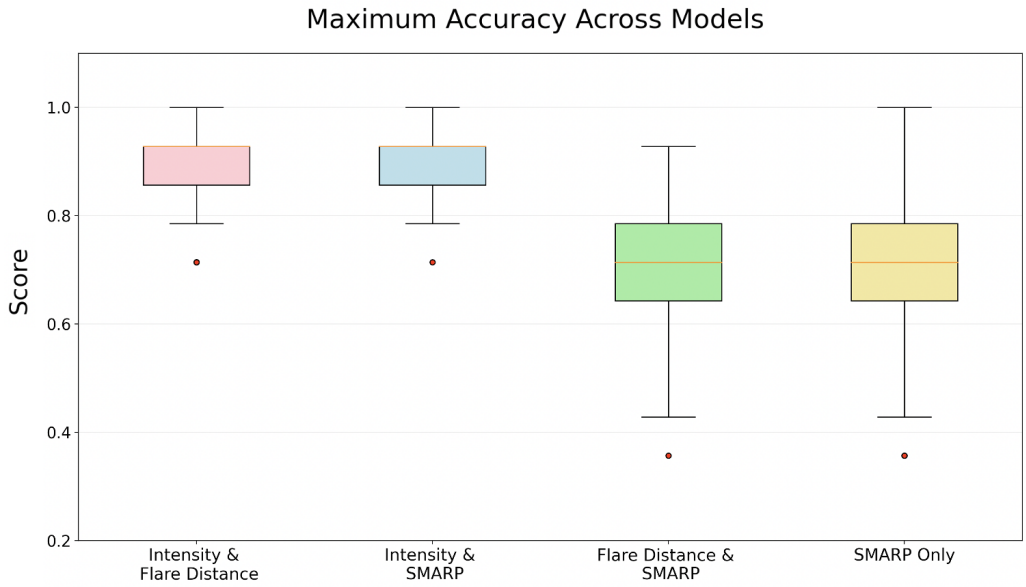

Figure 6. A cumulative box plot for the four main categories of predictor combinations outlined in the Appendix Tables. More specifically, the first plot (pink) corresponds to row 1 in Table 3, the second (blue) to rows 2-5, the third (green) to 6-10 and the fourth (yellow) corresponds to Table 2. The plot makes evident the superiority of the flare peak intensity over the SMARP data.

On the other hand, Papaioannou et al. (2018) perform a principal component analysis (PCA) on a set of six solar variables obtained from GOES and LASCO in order to calculate a decision boundary for their logistic regression. They classify events as SEP versus non-SEP and achieve a maximum POD (TSS + POFD) of 77.78%. Based on flare prediction, the warning tool García-Rigo et al. (2016) present provides long-term warnings of possible SEP event occurrence with POD scores of up to 58.3%. Núñez (2011) presents a dual-model system called UMASEP that has a POD of all (well and poorly connected with flares) SEP events of 80.72%. The SMARP data in Figure 3 achieves a POD 78.8%, similar to the works of Papaioannou and Nunez. If intensity gets involved in our logistic regression model, we can achieve POD scores of up to 90%.

All the results we report are using a probability threshold $p_t = 0.5$ because it is the one that is going to yield the optimal results for our models as demonstrated in Figure 7. It is possible for works that use different models and datasets to slightly increase their prediction statistics as Anastasiadis et al. (2017) show in their work.

To demonstrate how the skill scores change when varying the decision threshold, we show how the combination of ARDIST and USFLUX (one of the best performing SMARP predictor combinations of Table 2) performs on the Logistic Regression model. On the left panel of Figure 7, a graph of the contingency table values is presented as the decision threshold T_d changes from 0.3 to 0.7 in 0.05 increments. On the right one, the three different skill scores are being calculated and their box plots are graphed in order to present the uncertainties of each different decision threshold. Different models show similar trends with this specific case when the decision threshold is varied.

Lastly, it is important to note that none of the skill scores presented in this work can be immediately compared to the ones in the cited literature because the underlying class ratio and the train and testing data is not the same. It is therefore impossible

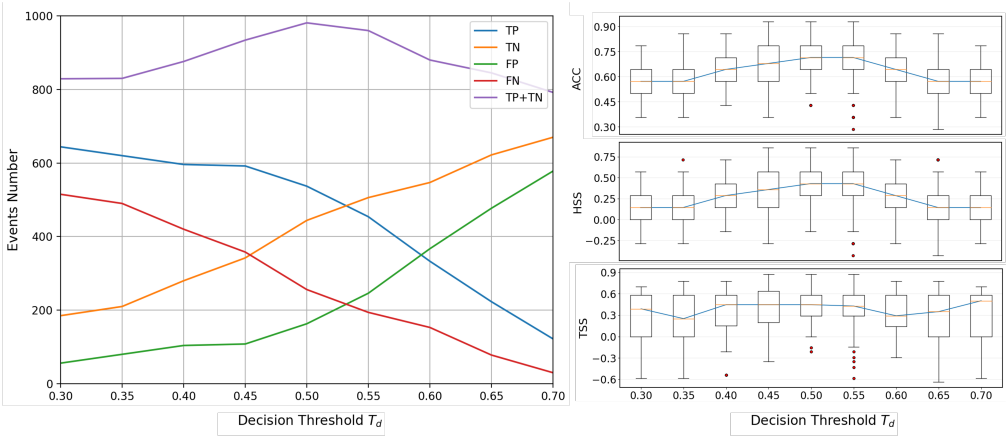

Figure 7. The left panel presents the contingency table values (TP, TN, FP and FN) along with the total number of correct predictions (TP+TN) for a nuber of different decision thresholds T_p. We observe that as the decision threshold increases, the number of false alarms (in green) increases while the number of missed events (in red) decreases. The total number of correct predictions is maximum where $T_d = 0.5$, therefore the results presented in the paper are calculated using this specific value. The right panel includes three different graphs for every one of the skill scores discussed in the Section 5 (ACC, TSS and HSS). The ACC and HSS mean values are at maximum when T_d is 0.50 and 0.55, while the TSS is mainly constant around 0.4 except a couple of decision threshold cases (0.35 and 0.6).

to do a fair comparison as SMARP is a newly published dataset and this is the first work that uses it for SEP prediction. Although we acknowledge that the comparisons with the literature should not be immediate, this is not a problem as the aim of this research is not to push the SEP prediction accuracy boundaries but rather to evaluate the SMARP dataset and its different predictors. This is the reason extra comparison runs using the GOES X-ray data are performed.

5.4 Connections to Operational Forecasting

Although the SMARP results presented above show that the dataset has some potential predictive capability, given its limitations (low resolution, limited positive events etc.), the SMARP data series could not be used alone to reliably forecast SEP events. Although the aim of the study is not to propose an operational forecasting apparatus, it is of interest to explore how the SMARP predictor models would perform if they were to be deployed. Three additional experiments (Experiment B, C and Operational in Table 5.4) are performed using the Support Vector Machine model of Figure 3 where the unsigned flux and the active region distance SMARPs are chosen to be the predictors. The differences between the experiments presented in Table 5.4 lie on the way the SMARP dataset is utilized during training and testing. The first experiment is the one discussed in Section 5.1 where a balanced number of positive and negative events is used during both training and testing. One way for handling imbalanced classes in SVMs is by introducing a hyperparameter C which determines the penalty for misclassification. In the last three experiments of Table 5.4, C is weighted in such a way (1:100 imbalance ratio) that it increases the penalty for misclassifying positive samples (the minority class) to prevent them from being overwhelmed by the negative ones (majority class).

In the second experiment (B), although during testing the same number (7) of positive and negative samples is used (balanced testing), the SVM is trained using the en-

Experiment	Training	Ratio	Testing	Ratio	ACC	TSS	HSS	POD	FAR
Original	Balance	1:1	Balance	1:1	0.70 ± 0.12	0.43 ± 0.25	0.39 ± 0.23	0.78	0.31
B	Imbalance	1:100	Balance	1:1	0.69 ± 0.12	0.42 ± 0.27	0.39 ± 0.25	0.78	0.33
C	Imbalance	1:100	Imbalance	1:10	0.52 ± 0.04	0.01 ± 0.01	0.01 ± 0.01	0.80	0.98
Operational	Imbalance	1:100	Unknown	1:X	0.52 ± 0.05	0.01 ± 0.01	0.01 ± 0.01	0.81	0.98

Table 1. The four different experiments performed using the USFLUX and the AR Distance predictors on a Third Degree Polynomial Support Vector Machine model. The ACC, TSS, HSS results are given in the format of mean ± standard deviation.

tire dataset mentioned in Section 2.2 comprised of 65 positive and 6,510 negative samples. In Experiments C and Opertational, we further mimic a forecasting apparatus in operational use in terms of having unbalanced samples: in Experiment C, the test set is made up of 70 negative samples 7 positive samples; whereas in Experiment Operational, the test set is randomly sampled from the full list of positive and negative samples. Again, for all the experiments, we follow the idea of the bootstrap procedure as described in Section 4.3 to assess the variability of the results, as given by the standard deviations of ACC, TSS, and HSS in Table 5.4. The results in Table 5.4 show that although the SMARP dataset has some ability to make distinctions between flares that produce SEPs and flares that do not, when put in an operational setting where the testing is performed in a vastly imbalanced set of samples, the forecasting model fails to produce meaningful results. It is important to note that for the models which imitate an operational forecaster, the Probability of Detection (POD) is higher than the more experimental models, but the increase in False Alarms results to low ACC, TSS and HSS scores. This is yet another proof of the initial hypothesis that although the SMARP dataset includes meaningful information which can be proven useful for SEP forecasting, it cannot be used by itself as a forecasting dataset.

6 Conclusions

To predict SEP events we use the newly published Space-Weather MDI Active Region Patches (SMARPs) dataset which includes observations of the solar magnetogram that were made during the active Solar Cycle 23. Point data selected from the SMARP time series is used on a variety of machine learning algorithms such as a different Support Vector Machines and Linear Regression models. The purpose of this study is to evaluate the power of this new data product for SEP forecast. Our results (Table 3 & 2) show that SMARP can accomplish this task as it can identify correctly 72% of the times whether an Active Region that produces a flare will lead to an SEP or not. Although the prediction results for the SMARP dataset are worse than the ones produced using the flare peak intensity and location, we demonstrate that not only SMARP data produces better results compared to earlier SEP prediction works, but it also provides a better leading time than other datasets.

The task of SEP prediction using SMARP data is subject to inherent limitations such as data uncertainties and a vastly unbalanced set of datasets which only includes a limited amount of positive events. To overcome these difficulties, a Monte Carlo method of random sampling, i.e. a bootstrap procedure, was employed to quantify of the results' uncertainties. It is important to note that the results presented in this paper should not be considered as an estimate of the accuracy that a prediction apparatus would yield if deployed but should rather be viewed as an effort to quantify and compare the prediction capability of the flare and SMARP predictors. In conclusion, although the SMARP dataset is constructed from the MDI data set, which includes only the line-of-sight component of the surface magnetic field at a relatively long 96-minute cadence, it can pro-

duce competitive prediction results for SEPs while providing a longer leading time than using Flare Predictors.

Acknowledgments

This work was supported by NASA DRIVE Science Center grant 80NSSC20K0600. The SEP event list we use in this work is documented in the NOAA Space Environment Service Center website (`https://www.ngdc.noaa.gov/stp/satellite/goes/doc/SPE.txt`), the SMARP data along with full-disk Helioseismic and Magnetic Imager (HMI) magnetograms is available on the Joint Science Operations Center database at `http://jsoc.stanford.edu/` and the NOAA solar X-ray flare dataset can be found at `https://www.ngdc.noaa.gov/stp/solar/solarflares.html`. All codes and data are included in our Github repository at `https://github.com/skasapis/SEP_Prediction_Using_SMARP`.

Table 2. Maximum ACC, TSS and HSS Values for the SVM and Linear Models using SMARP Predictors

	SVMs				Linear Models		
SMARP Predictors	Linear	RBF	Polynomial 2	Polynomial 3	Logistic Reg.	Ridge	Score
	0.67 ± 0.12	0.67 ± 0.13	**0.70 ± 0.12**	**0.72 ± 0.12**	**0.70 ± 0.12**	**0.71 ± 0.12**	ACC
1. USFLUXL & ARDIST	0.39 ± 0.28	0.38 ± 0.28	0.43 ± 0.25	0.47 ± 0.24	0.43 ± 0.23	0.47 ± 0.25	TSS
	0.34 ± 0.24	0.34 ± 0.25	0.39 ± 0.23	0.44 ± 0.25	0.40 ± 0.24	0.42 ± 0.24	HSS
	0.65 ± 0.11	0.67 ± 0.12	0.65 ± 0.12	0.67 ± 0.12	0.69 ± 0.11	0.65 ± 0.12	ACC
2. USFLUXL & ARAREA	0.35 ± 0.27	0.38 ± 0.28	0.34 ± 0.27	0.36 ± 0.26	0.37 ± 0.23	0.30 ± 0.27	TSS
	0.30 ± 0.23	0.35 ± 0.24	0.30 ± 0.24	0.33 ± 0.24	0.34 ± 0.22	0.27 ± 0.23	HSS
	0.69 ± 0.11	0.69 ± 0.11	0.52 ± 0.04	0.51 ± 0.03	0.67 ± 0.12	**0.70 ± 0.12**	ACC
3. ARDIST & ARAREA	0.42 ± 0.25	0.42 ± 0.23	0.15 ± 0.28	0.10 ± 0.22	0.36 ± 0.25	0.42 ± 0.26	TSS
	0.37 ± 0.23	0.38 ± 0.22	0.05 ± 0.09	0.03 ± 0.06	0.34 ± 0.24	0.40 ± 0.25	HSS
	0.65 ± 0.13	0.68 ± 0.10	0.58 ± 0.11	0.60 ± 0.11	0.67 ± 0.11	0.66 ± 0.12	ACC
4. ARDIST & RVALUE	0.33 ± 0.28	0.38 ± 0.22	0.18 ± 0.29	0.26 ± 0.29	0.36 ± 0.23	0.35 ± 0.25	TSS
	0.31 ± 0.26	0.35 ± 0.20	0.15 ± 0.22	0.21 ± 0.23	0.34 ± 0.21	0.33 ± 0.23	HSS
	0.67 ± 0.13	0.68 ± 0.11	**0.70 ± 0.11**	0.67 ± 0.10	**0.70 ± 0.13**	0.69 ± 0.10	ACC
5. USFLUXL, ARDIST	0.36 ± 0.28	0.38 ± 0.24	0.42 ± 0.23	0.37 ± 0.21	0.42 ± 0.27	0.41 ± 0.22	TSS
& ARAREA	0.34 ± 0.26	0.35 ± 0.22	0.39 ± 0.22	0.34 ± 0.19	0.39 ± 0.26	0.38 ± 0.21	HSS
	0.68 ± 0.12	0.68 ± 0.13	0.66 ± 0.13	0.69 ± 0.12	0.67 ± 0.11	0.69 ± 0.13	ACC
6. All SMARP Predictors	0.40 ± 0.24	0.32 ± 0.27	0.35 ± 0.28	0.42 ± 0.25	0.36 ± 0.24	0.40 ± 0.27	TSS
	0.37 ± 0.23	0.35 ± 0.25	0.33 ± 0.27	0.39 ± 0.24	0.33 ± 0.23	0.38 ± 0.26	HSS

Note. The ACC values $\geq$ 0.70 are marked in bolt. In green and red are marked the higher and lower accuracy values respectively.

Appendix

Table 3. Maximum ACC, TSS and HSS Values for the SVM and Linear Models using Different Predictors

Flare Predictors	SVMs				Linear Models		
	Linear	RBF	Polynomial 2	Polynomial 3	Logistic Reg.	Ridge	
1. Intensity & Flare Distance	0.90 ± 0.08	**0.91 ± 0.07**	0.90 ± 0.08	0.90 ± 0.08	0.90 ± 0.08	**0.91 ± 0.07**	ACC
	0.82 ± 0.16	0.84 ± 0.13	0.82 ± 0.15	0.80 ± 0.16	0.83 ± 0.16	0.84 ± 0.12	TSS
	0.80 ± 0.17	0.82 ± 0.14	0.80 ± 0.16	0.78 ± 0.17	0.80 ± 0.17	0.82 ± 0.14	HSS
SMARP & Flare Predictors							
2. Flare Distance & USFLUXL	0.71 ± 0.10	0.68 ± 0.13	0.70 ± 0.11	0.70 ± 0.11	0.71 ± 0.11	0.71 ± 0.12	ACC
	0.46 ± 0.20	0.39 ± 0.28	0.44 ± 0.24	0.42 ± 0.24	0.45 ± 0.22	0.46 ± 0.25	TSS
	0.42 ± 0.20	0.36 ± 0.26	0.40 ± 0.23	0.39 ± 0.23	0.42 ± 0.21	0.41 ± 0.23	HSS
3. Flare Distance & RVALUE	0.67 ± 0.14	0.69 ± 0.13	0.61 ± 0.14	0.61 ± 0.12	0.69 ± 0.12	0.70 ± 0.10	ACC
	0.36 ± 0.29	0.41 ± 0.27	0.25 ± 0.32	0.30 ± 0.32	0.41 ± 0.25	0.43 ± 0.22	TSS
	0.34 ± 0.27	0.38 ± 0.26	0.22 ± 0.27	0.23 ± 0.23	0.38 ± 0.24	0.40 ± 0.21	HSS
4. Flare Distance & ARAREA	0.66 ± 0.13	0.68 ± 0.11	0.60 ± 0.09	0.62 ± 0.10	0.70 ± 0.12	0.69 ± 0.11	ACC
	0.35 ± 0.27	0.40 ± 0.25	0.36 ± 0.30	0.39 ± 0.29	0.43 ± 0.24	0.42 ± 0.24	TSS
	0.32 ± 0.26	0.36 ± 0.22	0.19 ± 0.18	0.24 ± 0.21	0.41 ± 0.23	0.39 ± 0.22	HSS
5. Flare Distance, USFLUXL & ARAREA	0.69 ± 0.13	0.69 ± 0.13	0.67 ± 0.12	0.66 ± 0.11	0.69 ± 0.13	0.69 ± 0.11	ACC
	0.41 ± 0.27	0.42 ± 0.27	0.35 ± 0.30	0.36 ± 0.25	0.35 ± 0.26	0.41 ± 0.24	TSS
	0.38 ± 0.26	0.37 ± 0.26	0.32 ± 0.28	0.32 ± 0.23	0.39 ± 0.25	0.38 ± 0.22	HSS
6. Intensity & USFLUXL	0.88 ± 0.09	0.89 ± 0.07	0.90 ± 0.08	0.90 ± 0.08	0.88 ± 0.09	0.89 ± 0.09	ACC
	0.80 ± 0.16	0.80 ± 0.13	0.82 ± 0.14	0.80 ± 0.17	0.78 ± 0.17	0.79 ± 0.17	TSS
	0.77 ± 0.18	0.77 ± 0.15	0.80 ± 0.15	0.79 ± 0.18	0.76 ± 0.18	0.77 ± 0.17	HSS
7. Intensity & RVALUE	0.89 ± 0.08	**0.91 ± 0.07**	0.91 ± 0.08	0.89 ± 0.09	**0.91 ± 0.07**	0.90 ± 0.08	ACC
	0.79 ± 0.16	0.83 ± 0.14	0.83 ± 0.15	0.80 ± 0.16	0.84 ± 0.13	0.83 ± 0.15	TSS
	0.77 ± 0.17	0.82 ± 0.15	0.81 ± 0.16	0.78 ± 0.17	0.81 ± 0.14	0.80 ± 0.16	HSS
8. Intensity & ARDIST	**0.91 ± 0.07**	**0.91 ± 0.07**	**0.91 ± 0.07**	0.90 ± 0.07	**0.92 ± 0.07**	**0.91 ± 0.07**	ACC
	0.84 ± 0.13	0.84 ± 0.13	0.83 ± 0.14	0.83 ± 0.13	0.86 ± 0.13	0.84 ± 0.14	TSS
	0.82 ± 0.14	0.81 ± 0.14	0.81 ± 0.15	0.81 ± 0.14	0.84 ± 0.15	0.82 ± 0.15	HSS
9. Intensity, USFLUXL & ARDIST	**0.91 ± 0.07**	0.91 ± 0.08	0.90 ± 0.08	**0.91 ± 0.07**	**0.92 ± 0.08**	0.90 ± 0.08	ACC
	0.83 ± 0.14	0.84 ± 0.15	0.82 ± 0.15	0.83 ± 0.13	0.85 ± 0.15	0.81 ± 0.16	TSS
	0.82 ± 0.15	0.82 ± 0.16	0.80 ± 0.16	0.82 ± 0.14	0.83 ± 0.16	0.80 ± 0.17	HSS
10. Intensity & MEANGBL	0.90 ± 0.08	**0.91 ± 0.07**	0.91 ± 0.08	0.90 - 0.08	0.91 ± 0.08	0.90 ± 0.08	ACC
	0.82 ± 0.15	0.85 ± 0.13	0.84 ± 0.15	0.82 ± 0.15	0.84 ± 0.14	0.82 ± 0.16	TSS
	0.80 ± 0.17	0.83 ± 0.14	0.83 ± 0.15	0.80 ± 0.16	0.82 ± 0.15	0.80 ± 0.16	HSS

Note. The ACC values $\geq$ 0.91 with standard deviation $\leq$ 0.07 are marked in bold. In green and red are marked the higher and lower accuracy values respectively for each one of the three predictor groups.

References

Amari, S.-i., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*(6), 783–789.

Anastasiadis, A., Papaioannou, A., Sandberg, I., Georgoulis, M., Tziotziou, K., Kouloumvakos, A., & Jiggens, P. (2017). Predicting flares and solar energetic particle events: The forspef tool. *Solar Physics*, *292*(9), 1–21.

Bain, H., Brea, P., & Adamson, E. (2018). Using machine learning techniques to forecast solar energetic particles. In *Agu fall meeting abstracts* (Vol. 2018, pp. SM31D–3530).

Balch, C. C. (2008). Updated verification of the space weather prediction center's solar energetic particle prediction model. *Space Weather*, *6*(1).

Barnes, W. T., Bobra, M. G., Christe, S. D., Freij, N., Hayes, L. A., Ireland, J., ... others (2020). The sunpy project: Open source development and status of the version 1.0 core package. *The Astrophysical Journal*, *890*(1), 68.

Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. (2021). Smarps and sharps: Two solar cycles of active region data.

Brueckner, G., Howard, R., Koomen, M., Korendyke, C., Michels, D., Moses, J., ... others (1995). The large angle spectroscopic coronagraph (lasco). In *The soho mission* (pp. 357–402). Springer.

Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., DuFumier, B., Zhou, T., ... Gombosi, T. I. (2019). Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters. *Space Weather*, *17*(10), 1404–1426.

Choudhary, D. P., Gosain, S., Gopalswamy, N., Manoharan, P., Chandra, R., Uddin, W., ... others (2013). Flux emergence, flux imbalance, magnetic free energy and solar flares. *Advances in Space Research*, *52*(8), 1561–1566.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1-26.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

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 Physics*, *293*(2), 1–42.

Forbush, S. E. (1946). Three unusual cosmic-ray increases possibly due to charged particles from the sun. *Physical Review*, *70*(9-10), 771.

García-Rigo, A., Núñez, M., Qahwaji, R., Ashamari, O., Jiggens, P., Pérez, G., ... Hilgers, A. (2016). *Prediction and warning system of sep events and solar flares for risk estimation in space launch operations.* EDP Sciences.

Gurman, J. B., Withbroe, G. L., & Harvey, J. W. (1974). A comparison of euv spectroheliograms and photospheric magnetograms. *Solar Physics*, *34*(1), 105–111.

Hanssen, A., & Kuipers, W. (1965). *On the relationship between the frequency of rain and various meteorological parameters.* Koninklijk Nederlands Meteorologisch Instituut.

Heidke, P. (1926). Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geografiska Annaler*, *8*(4), 301–349.

Horne, R., Glauert, S., Meredith, N., Boscher, D., Maget, V., Heynderickx, D., & Pitchford, D. (2013). Space weather impacts on satellites and forecasting the earth's electron radiation belts with spacecast. *Space Weather*, *11*(4), 169–186.

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*(2), 415–425.

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. *The Astrophysical Journal*, *861*(2), 128.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, *68*(6), 540.

Laurenza, M., Cliver, E., Hewitt, J., Storini, M., Ling, A., Balch, C., & Kaiser, M. (2009). A technique for short-term warning of solar energetic particle events based on flare location, flare size, and evidence of particle escape. *Space Weather*, *7*(4).

Lavasa, E., Giannopoulos, G., Papaioannou, A., Anastasiadis, A., Daglis, I., Aran, A., & Pacheco, D. (2021). Assessing the predictability of solar energetic particles with the use of machine learning techniques.

Lopez, R. E., Baker, D. N., & Allen, J. (2004). Sun unleashes halloween storm. *Eos, Transactions American Geophysical Union*, *85*(11), 105–108.

McCracken, K., & Ness, N. (1966). The collimation of cosmic rays by the interplanetary magnetic field. *Journal of Geophysical Research*, *71*(13), 3315–3318.

Mumford, S. J., Christe, S., Pérez-Suárez, D., Ireland, J., Shih, A. Y., Inglis, A. R., ... others (2015). Sunpypython for solar physics. *Computational Science & Discovery*, *8*(1), 014009.

Núñez, M. (2011). Predicting solar energetic proton events (e¿ 10 mev). *Space Weather*, *9*(7).

Papaioannou, A., Anastasiadis, A., Kouloumvakos, A., Paassilta, M., Vainio, R., Valtonen, E., ... Abunin, A. (2018). Nowcasting solar energetic particle events using principal component analysis. *Solar Physics*, *293*(7), 1–23.

Papaioannou, A., Sandberg, I., Anastasiadis, A., Kouloumvakos, A., Georgoulis, M. K., Tziotziou, K., ... Hilgers, A. (2016). Solar flares, coronal mass ejections and solar energetic particle event characteristics. *Journal of Space Weather and Space Climate*, *6*, A42.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Reames, D. V. (2013). The two sources of solar energetic particles. *Space Science Reviews*, *175*(1-4), 53–92.

Richardson, I., Mays, M., & Thompson, B. (2018). Prediction of solar energetic particle event peak proton intensity using a simple algorithm based on cme speed and direction and observations of associated solar phenomena. *Space Weather*, *16*(11), 1862–1881.

Rodriguez, J., Krosschell, J., & Green, J. (2014). Intercalibration of goes 8–15 solar proton detectors. *Space Weather*, *12*(1), 92–109.

Sadykov, V., Kosovichev, A., Kitiashvili, I., Oria, V., Nita, G. M., Illarionov, E., ... Ali, A. (2021). Prediction of solar proton events with machine learning: Comparison with operational forecasts and" all-clear" perspectives. *arXiv preprint arXiv:2107.03911*.

Saputro, D. R. S., & Widyaningsih, P. (2017). Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). In *Aip conference proceedings* (Vol. 1868, p. 040009).

Scherrer, P. H., Bogart, R. S., Bush, R., Hoeksema, J.-A., Kosovichev, A., Schou, J., ... others (1995). The solar oscillations investigationmichelson doppler imager. In *The soho mission* (pp. 129–188). Springer.

Schou, J., Scherrer, P. H., Bush, R. I., Wachter, R., Couvidat, S., Rabello-Soares, M. C., ... others (2012). Design and ground calibration of the helioseismic and magnetic imager (hmi) instrument on the solar dynamics observatory (sdo). *Solar Physics*, *275*(1), 229–259.

Schrijver, C., Beer, J., Baltensperger, U., Cliver, E., Güdel, M., Hudson, H., ... others (2012). Estimating the frequency of extremely energetic solar events, based on solar, stellar, lunar, and terrestrial records. *Journal of Geophysical Research: Space Physics*, *117*(A8).

Shao, Y.-H., Chen, W.-J., & Deng, N.-Y. (2014). Nonparallel hyperplane support vector machine for binary classification problems. *Information Sciences*, *263*, 22–35.

Shea, M., & Smart, D. (1995). History of solar proton event observations. *Nuclear Physics B-Proceedings Supplements*, *39*(1), 16–25.

Ugarte-Urra, I., Upton, L., Warren, H. P., & Hathaway, D. H. (2015). Magnetic flux transport and the long-term evolution of solar active regions. *The Astrophysical Journal*, *815*(2), 90.

van Driel-Gesztelyi, L., Démoulin, P., Mandrini, C. H., Harra, L., & Klimchuk, J. (2003). The long-term evolution of ar 7978: The scalings of the coronal plasma parameters with the mean photospheric magnetic field. *The Astrophysical Journal*, *586*(1), 579.

van Driel-Gesztelyi, L., & Green, L. M. (2015). Evolution of active regions. *Living Reviews in Solar Physics*, *12*(1), 1–98.

Vapnik, V. (1998). The support vector method of function estimation. In *Nonlinear modeling* (pp. 55–85). Springer.

Webb, D. F., & Allen, J. H. (2004). Spacecraft and ground anomalies related to the october-november 2003 solar activity. *Space Weather*, *2*(3).

Wild, J., Smerd, S., & Weiss, A. (1963). Solar bursts. *Annual Review of Astronomy and Astrophysics*, *1*, 291.