The Astrophysical Journal, 974:131 (13pp), 2024 October 10 https://doi.org/10.3847/1538-4357/ad6f0e

# Forecasting Solar Energetic Particle Events During Solar Cycles 23 and 24 Using Interpretable Machine Learning

Spiridon Kasapis[1], Irina N. Kitiashvili[1], Paul Kosovich[2], Alexander G. Kosovichev[1,2], Viacheslav M. Sadykov[3], Patrick O'Keefe[2], and Vincent Wang[4]

[1] NASA Ames Research Center, Moffett Field, CA, USA
[2] New Jersey Institute of Technology, Newark, NJ, USA
[3] Georgia State University, Atlanta, GA, USA
[4] Purdue University, West Lafayette, IN, USA

*Received 2024 February 16; revised 2024 May 17; accepted 2024 May 24; published 2024 October 9*

## Abstract

The prediction of solar energetic particle (SEP) events garners increasing interest as space missions extend beyond Earth's protective magnetosphere. These events, which are, in most cases, products of magnetic-reconnection-driven processes during solar flares or fast coronal-mass-ejection-driven shock waves, pose significant radiation hazards to aviation, space-based electronics, and particularly space exploration. In this work, we utilize the recently developed data set that combines the Solar Dynamics Observatory/Space-weather Helioseismic and Magnetic Imager Active Region Patches and the Solar and Heliospheric Observatory/Space-weather Michelson Doppler Imager Active Region Patches. We employ a suite of machine learning strategies, including support vector machines (SVMs) and regression models, to evaluate the predictive potential of this new data product for a forecast of post-solar flare SEP events. Our study indicates that despite the augmented volume of data, the prediction accuracy reaches $0.7 \pm 0.1$ (experimental setting), which aligns with but does not exceed these published benchmarks. A linear SVM model with training and testing configurations that mimic an operational setting (positive–negative imbalance) reveals a slight increase ($+0.04 \pm 0.05$) in the accuracy of a 14 hr SEP forecast compared to previous studies. This outcome emphasizes the imperative for more sophisticated, physics-informed models to better understand the underlying processes leading to SEP events.

*Unified Astronomy Thesaurus concepts:* Solar energetic particles (1491); Solar storm (1526); Solar particle emission (1517); Support vector machine (1936); Regression (1914); Linear regression (1945)

## 1. Introduction

Solar energetic particle (SEP) events are one of the manifestations of solar activity that may significantly impact the conditions of the space environment. For example, the large solar particle event of 2017 September emphasized a significant surge in the charged and neutral particle flux that was able to reach Mars' surface (C. Zeitlin et al. 2018). While the doses from this specific event were below NASA's stipulated radiation exposure limits for astronauts, the risk for future explorers is evident. This concern becomes particularly relevant in scenarios where human explorers might be far from their habitats on other celestial bodies, with the onset of an event leaving them vulnerable to enhanced radiation doses. Therefore, forecasting and predicting SEP events is paramount.

SEP events vary in intensity, spanning from suprathermal (few keV) up to relativistic (few GeV) energies, and are accelerated near the Sun either by magnetic-reconnection-driven processes during solar flares or by fast coronal mass ejections. SEP events are categorized as impulsive and gradual (M. Desai & J. Giacalone 2016). The large gradual SEP events are of special interest because they are accompanied by high-energy protons ($\geqslant$10 MeV) that pose serious radiation threats to aviation and space operations. While these events are intriguing from a scientific viewpoint, they have not always been consistently detected. Therefore, while over 200 major solar proton events were observed over the last 40 yr, discerning a pattern between these occurrences and the solar cycle remains elusive, adding to the difficulty in inferring the number and intensity of major solar proton events that can affect Earth (M. Shea & D. Smart 1990).

Many models have been developed to predict SEP events (K. Whitman et al. 2022). These models vary in their intricacies, ranging from physics-based (e.g., M. Zhang & L. Zhao 2017) to empirical (e.g., A. Anastasiadis et al. 2017; M. K. Georgoulis et al. 2021) and probabilistic (e.g., A. Papaioannou et al. 2022). Lately, the embrace of machine learning (ML) and hybrid model approaches signals a multidimensional approach toward tackling this issue. For instance, the Space Radiation Intelligence System merges preeruptive and posteruptive metadata, offering predictions for various solar-driven events, including SEPs (A. Engell et al. 2017).

Leveraging vast data sets, ML models are being trained to predict solar proton events (SPEs) by considering diverse parameters, such as the magnetic field characteristics of solar active regions (ARs), the preceding soft X-ray and proton fluxes, and the statistics of solar radio bursts. For example, ML models have demonstrated predictive capabilities superior to conventional SWPC NOAA forecasts, indicating the possibility of developing robust "all-clear" SPE forecasts (e.g., V. Sadykov et al. 2021). A. Ji et al. (2020) have demonstrated that through data collection and intricate predictive model building, a time series classification apparatus can offer more precise forecasts compared to baseline models by tuning model hyperparameters. It is also critical to address the challenges posed by imbalanced data sets in forecasting SPEs. As highlighted by M. Stumpo et al. (2021), it is important for machine learning techniques to reinterpret traditional statistical SPE forecasting models (M. Laurenza et al. 2009), demonstrating that there are validation strategies that enhance the forecast accuracy by balancing class distributions. Similarly, E. Lavasa et al. (2021) tackle SEP event prediction as an imbalanced binary classification problem, emphasizing the necessity of class-dependent penalization to accurately reflect the rarity of SEP events that we encounter in real-world applications.

In this study, we build upon the current state of SEP modeling, utilizing a data set that spans two solar cycles. This extended data set offers a new perspective on the SEP prediction problem, aiming to improve the accuracy and reliability of the existing models. To do this, we take advantage of our previous developments (S. Kasapis et al. 2022) and a new data set that covers two solar cycles, 23 and 24 (P. A. Kosovich et al. 2024), briefly described in Section 2. Because the prediction method is based on the known connection between SEP events and solar flares (G.-M. Le & X.-F. Zhang 2017), Section 3 outlines the process of matching these space-weather phenomena that follow up statistical analysis presented in Section 4. Finally, in Sections 5 and 6, we present the results of this study and the ML models, while in Section 7 we summarize the results that span the two solar cycle observations.

## 2. Solar Cycles 23 and 24 Active Regions Data Set

It is well known that not all ARs exhibit eruptive activity, and only a few of the ones that do are a source of SEP events (H. Cane et al. 2010). Over Solar Cycle 23, although the number of ARs and flares that erupted within them are on the order of thousands, only 93 SEPs were recorded by the NASA Space Radiation Analysis Group[5] (SRAG). This creates a significant statistical disproportion expectation of SEP events (so-called class-imbalance problem) when the forecast is based on the thousands of flares, and as a result, models often tend to miss them. Therefore, increasing the size of the data set suitable for ML training, by combining two solar cycles worth of data (168 SEPs in Solar Cycles 23 and 24), may improve the robustness of the SEP events forecast. Because most SEP events are associated with flares that erupted in ARs, it is natural to utilize the available Space-weather Helioseismic and Magnetic Imager (MDI) Active Region Patches (SMARP; M. Bobra 2017) and Space-weather Michelson Doppler Imager (HMI) Active Region Patches (SHARP; M. Bobra et al. 2011). The SMARP and SHARP data sets include maps of automatically tracked ARs extracted from full-disk magnetograms. These AR patches and their parameters have been used for a number of solar flare prediction studies (e.g., M. G. Bobra & S. Couvidat 2015; I. Kontogiannis et al. 2017; J. Wang et al. 2023). The vast majority of previous studies did not take advantage of either SMARP or SHARP. Training a machine learning model with the SMARP data set demonstrated a potential to predict SEP events (S. Kasapis et al. 2022), where training was based on five predictors: the total unsigned magnetic flux, the vertical field gradient, the unsigned flux $R$ near the polarity inversion lines (C. J. Schrijver 2007), the angular distance between an AR and Earth's magnetic foot point (A. Ippolito et al. 2005), and the area of an AR.

We use a time series of summary parameters (keywords) of solar ARs from both the SMARP and SHARP data series (M. Bobra et al. 2011; M. Bobra 2017; M. G. Bobra et al. 2021), which were combined into a single consistent data set (P. A. Kosovich et al. 2024). This merged data set keeps the original cadence of the data products: 96 minutes for MDI and 12 minutes for HMI observations. The SMARP/SHARP data set includes a continuous set of 21 keywords (Table 1), representing homogeneous observations of AR patches from the Solar and Heliospheric Observatory (SOHO) MDI and Solar Dynamics Observatory (SDO) HMI for the period between 1996 April 4 and 2022 December 30. Because MDI did not observe the transverse magnetic field, we computed the vertical components of the unsigned flux and the mean magnetic field gradient from the line-of-sight parameters assuming that the magnetic field of ARs is predominantly radial. For both SHARP and SMARP data sets, the R_VALUE was recomputed by finding the common antilogarithm of the original R_VALUE, and SMARP MEANGBL units were converted from Gauss pixel$^{-1}$ to Gauss Mm$^{-1}$. Filtering was applied to exclude low-quality observables and any records corresponding to the Stonyhurst coordinates beyond ±65° longitude to mitigate the foreshortening effect near the solar limb. Six SMARP parameters (USFLUXL, MEANGBL, R_VALUE, CMASKL, MEANGBZ, and USFLUXZ) were rescaled by applying total least squares fitting, and the two data sets were merged on 2010 May 1, 00:00:00 TAI. For a more thorough overview of filtering, rescaling, and merging of the SHARP and SMARP keywords, we refer readers to P. A. Kosovich et al. (2024).

To evaluate the predictive capabilities of an ML-driven SEP analysis, we use parameters of the flaring ARs as precursors, which are stored as the SMARP/SHARP parameters (Table 1). In our research, we use eight of these parameters (marked bold in Table 1). Five physical parameters (R_VALUE, MEANGBL, MEANGBZ, USFLUXL, USFLUXZ) are used for the model's training and evaluation. These keywords (so-called "predictors" in machine learning) are used as input to an ML model. Spatial parameters CRLT_OBS and CRLN_OBS are also involved in training and evaluation as they define the magnetic connectivity of ARs to the Earth (A. Ippolito et al. 2005), which we also use as a predictor. The time variable, T_OBS, allows us to pick the correct SMARP/SHARP data records associated with an SEP. More information about how T_OBS is used and the selection of the appropriate prediction windows for this research can be found in the following two sections.

## 3. Coupling of SEP Events and Solar Flares

Most SEP events are associated with solar eruptive activity; therefore the NOAA solar X-ray flare catalog[6] and the list of SEP events provided by the NASA SRAG are used to identify the SMARP/SHARP data points (selected from the AR time series) that will train our ML models. Note here that the NASA SRAG SEP list contains fewer SEP events than other data sets (S. Rotti et al. 2022) as only SEPs that can be connected with certainty to flare events are included. The flare catalog contains 4238 flares recorded from 1996 April 22 to 2022 December 30, the period between SOHO's first record and the SHARP/SMARP processed data availability at the time this research was conducted. This catalog only contains flares for which all

[5] https://srag.jsc.nasa.gov/

[6] https://www.ngdc.noaa.gov/stp/solar/solarflares.html

**Table 1**
List of Merged SMARP/SHARP Keywords (P. A. Kosovich et al. 2024)

| Keyword | Description | Example |
|---|---|---|
| DBINDEX | Database Index | "mdi.smarp_cea_96m..." |
| T_OBS | Time of observation | 11/4/97 4:47 |
| UNIX_TIME | T_OBS expressed as time elapsed since Unix epoch (days) | 10169.19965 |
| ARPNUM | Number of AR patch | 581 |
| NOAA_AR | NOAA AR number that best matches ARPNUM | 8100 |
| NOAA_ARS | List of all NOAA ARs matching this ARPNUM | 8100 |
| **USFLUXL** | **Total line-of-sight unsigned flux (Maxwells)** | **3.18E+22** |
| **R_VALUE** | **Unsigned flux *R* near polarity inversion lines (Maxwells)** | **251995.71** |
| **MEANGBL** | **Mean value of line-of-sight field gradient (Gauss Mm$^{-1}$)** | **103.10** |
| **USFLUXZ** | **Vertical component of the total unsigned flux (Maxwells)** | **120.44** |
| **MEANGBZ** | **Mean value of the vertical field gradient (Gauss Mm$^{-1}$)** | **6.75E+22** |
| LAT_FWT | Stonyhurst latitude of flux-weighted center of active pixels (degrees) | −19.58 |
| **CRLT_OBS** | **Carrington latitude of the observer (degrees)** | **4.02** |
| LON_FWT | Stonyhurst longitude of flux-weighted center of active pixels (degrees) | 27.23 |
| **CRLN_OBS** | **Carrington longitude of the observer (degrees)** | **324.65** |
| CAR_ROT | Carrington rotation number of CRLN_OBS | 1929 |
| CMASKL | Cylindrical equal-area pixels in the AR | 182848.46 |
| CDELT1 | Map scale in degrees per pixel | 0.12 |
| DSUN_OBS | Distance from SOHO/SDO to the center of the Sun (meters) | 146808723630.28 |
| RSUN_OBS | Observed angular radius of the Sun (arcseconds) | 977.876787 |
| QUALITY | Quality index | 512 |

**Note.** In bold are the keywords used as predictors in this study, and as examples, the values that describe AR8100 on 1997 November 4, 04:47:00, are given. The ARPNUM is called "TARPNUM" in the SMARP data set and "HARPNUM" in the SHARP data set, but were renamed here for consistency.

information related to their location and intensity/class is available and only flares produced in ARs that are included in the SHARP/SMARP data set. Only flares of the C, M, and X classes are kept because the weaker flares do not produce SEP events (A. Papaioannou et al. 2016; H. Cane et al. 2010), allowing us to mitigate the class-imbalance problem. Thus, the total number of flares was reduced to 3421. The SEP list includes 168 events recorded from 1997 November 4 to 2023 May 9. We excluded 14 SEP events because they occurred on the far side of the Sun or they were not associated with flares. Our analysis does not include any SEP events on the limb or far side because of the absence of reliable observations. The SEP list provided by NASA SRAG contains 154 SEP–flare couples (SEPs associated with flares). The NOAA list includes only 115 flares that can be verified (matched) with these SEP–flare couples.

Similar to S. Kasapis et al. (2022), we define flares associated with SEP as a positive event and the flares that did not produce an SEP event as a negative event. Due to data gaps that exist in the merged SMARP/SHARP data set, an additional 5 positive and 60 negative of these SEP–flare couples had to be excluded. Therefore, the final count for the flares that will be used for SEP prediction is 110 positive and 3246 negative events. The SMARP/SHARP data points availability and selection are discussed in Section 4.

To perform the association of SEP events to solar flares, we use the NOAA flare keywords (*t_start*, *t_max*, *t_end*, *class*, *location*, and *AR*) in Table 2. Additional keywords were generated to reflect if a particular flare produced an SEP event or not (*SEP_Match* keyword), as well as additional characteristics of an eruptive event (*intensity*, *coords*, and *ang_dist* keywords). The *SEP_Match* keyword is particularly important in this project, as it is used as the target for our ML-ready data sets. This SEP–flare matching process has assigned 110 flares as positive (*SEP_Match* = "True") and 3246 as negative (*SEP_Match* = "False").

The flare *intensity* keyword in Table 2 is defined as $I = n \times f$, where $n$ corresponds to the intensity within the flare class (*class* keyword), and $f$ is the classification factor which depends on the class of a flare: $10^{-6}$ for "C" class flares, $10^{-5}$ for "M," and $10^{-4}$ for "X" class flares. The *coords* keyword is produced by converting the solar flare location coordinates from the NSEW format (*location*) to heliographic coordinates. The NSEW format is a string-based representation, where the first character indicates north (N) or south (S), followed by two digits representing the latitude. The fourth character represents east (E) or west (W),

**Table 2**
List of Keywords in the NOAA Flare Data Set (Top) and the Ones Created for the Convenience of Our Analysis (Bottom)

| **Default NOAA Flare Keywords** | | |
|---|---|---|
| **Keyword** | **Description** | **Example** |
| *t_start* | Flare start time | 1997-11-04 05:52:00 |
| *t_max* | Flare maximum intensity time | 1997-11-04 05:58:00 |
| *t_end* | Flare end time | 1997-11-04 06:02:00 |
| *class* | Flare class | X2.1 |
| *location* | Flare location (NESW format) | S14W33 |
| *AR* | NOAA AR number associated with flare | 8100 |
| **New Keywords** | | |
| **Keyword** | **Description** | **Example** |
| *SEP_Match* | SEP-producing flare (True or False) | True |
| *intensity* | Intensity (Watts per square meter) | 0.00021 |
| *coords* | Flare heliographic coordinates (degrees) | (−14, 33) |
| *ang_dist* | Angular distance from the magnetic foot point of the Earth (radians) | −0.32045 |

**Note.** Although the *SEP_Match* keyword is used for the ML models training, the *intensity*, *coords*, and *ang_dist* keywords are only used for demonstrating the relationship between flares and SEPs in Figure 1.

followed by two digits for the longitude. The heliographic coordinates are based on the Sun's rotation and are expressed as latitude and longitude in degrees, with N and W considered positive. Therefore, the NSEW coordinates "S14W33" would be converted to (−14, 33) in heliographic coordinates (Table 2, New Keywords). Finally, the *ang_dist* keyword represents the angular distance $\Delta\sigma$ between the AR and the magnetic foot point of Earth using the haversine formula (C. C. Robusto 1957), which is appropriate for calculating great-circle distances between two points on a sphere:

$$\Delta\sigma = \arccos(\sin\theta_1 \sin\theta_2 + \cos\theta_1 \cos\theta_2 \cos(\phi_1 - \phi_2)), \quad (1)$$

where ($\theta_1$, $\phi_1$) are the latitude and longitude of an AR, and ($\theta_2$, $\phi_2$) are the latitude and longitude of the magnetic foot point of Earth. The distance to the magnetic foot point of the Earth is critical for SEP prediction because it delineates the magnetic connectivity between solar flares and the Earth, serving as a pathway for SEPs to travel to our planet's vicinity (A. Ippolito et al. 2005). We assume that the magnetic foot point of Earth is at the position (0, 45) degrees in the Stonehurst heliographic coordinates, the location of the magnetic field line connecting the Sun and Earth on the source surface (2.5 solar radii from the solar center) where Parker's spiral originates. The coordinates are initially in degrees, but we convert them into radians to fit the Python trigonometric functions requirement. After the calculation, the angular distance $\Delta\sigma$ is returned in radians. An additional consideration is made for the ARs that are east of the magnetic foot point, in which case the distance is assigned a negative sign. Thus, using the SHARP/SMARP coordinates (CRLT_OBS and CRLN_OBS, Table 1) in Equation (1), we obtained the angular distance to the magnetic foot point of the Earth (ANG_DIST keyword) for the ARs in the SHARP/SMARP data set.

The *ang_dist* flare keyword is calculated using the flare location in Table 2, whereas, in our ML models, we will be using the AR location information from Table 1. The keyword values in Table 1 are the ones provided in the SMARP/SHARP data set, and we are going to be evaluating them in this research rather than the NOAA flare information. The solar flare keywords are not used to train the ML models as in an operational setting, at the moment of prediction, we do not have knowledge of the flare eruption and the physical parameters associated with it. The flare keywords are only used to assign targets (SEP or not) to the SMARP/SHARP data points and to shed light on the relationship between flares and the SEPs they produced. For example, the density histogram on the left panel of Figure 1 shows much less overlap between the positive (green) and negative (red) flares compared to the histograms on the right panel, indicating higher predictive capabilities of magnetic flux.

Although Figure 1 demonstrates that *intensity* is a relatively reliable predictor of SEP events, the right panel shows that the angular distance is not, but still carries some information related to SEP production. As expected, flares with low values of angular distance (flares that occurred close to the Earth–Sun magnetic connection point) are more likely to produce SEPs that can be observed at Earth, compared to flares that erupted far away from the magnetic foot point (Figure 1, right panel, angular distance values $\leqslant -1$). It is important to note, that such keywords can be used for SEP prediction only if a flare prediction model is able to infer the flare intensity and position on the solar disk (Z. Jiao et al. 2020; A. Chen et al. 2021). Although this work is limited to the predictors that the SHARP/SMARP data set offers, it is important to note that other SEP prediction and characterization studies have used a variety of different observables as inputs, such as the radio burst flux (S. W. Kahler & A. G. Ling 2017), the soft X-ray (SXR) flux, the SXR duration, the SXR rise time, and other (E. Lavasa et al. 2021; K. Whitman et al. 2022; P. M. O'Keefe et al. 2023).

## 4. SHARP/SMARP Data Selection and Creation of ML-ready Data Sets

To predict the SEP events, we consider the time evolution of an AR and its properties during the start time of a corresponding solar flare. In this work, we use 3869 ARs from the SMARP/SHARP data set (P. A. Kosovich et al. 2024), 872 of which produced a flare. Because the eruptive activity of the Sun is the primary source of the SEP events, we consider only flaring ARs (such as the two example ARs in Figure 2). Thus, we combine information available from the three data sources: (1) properties of flaring ARs from the SHARP/SMARP data set, (2) the NOAA catalog of solar flares, and (3) the SWPC catalog of SEP events.

After labeling every flare in the NOAA flare list (using the SEP_Match keyword) based on whether it is associated with an SEP (Section 3), we proceed with selecting the appropriate SHARP/SMARP data (Section 4) used to train various ML models. Figure 2 illustrates these two distinct processes of creating different data sets used in this paper (flare and SEP coupling, SMARP/SHARP data selection).

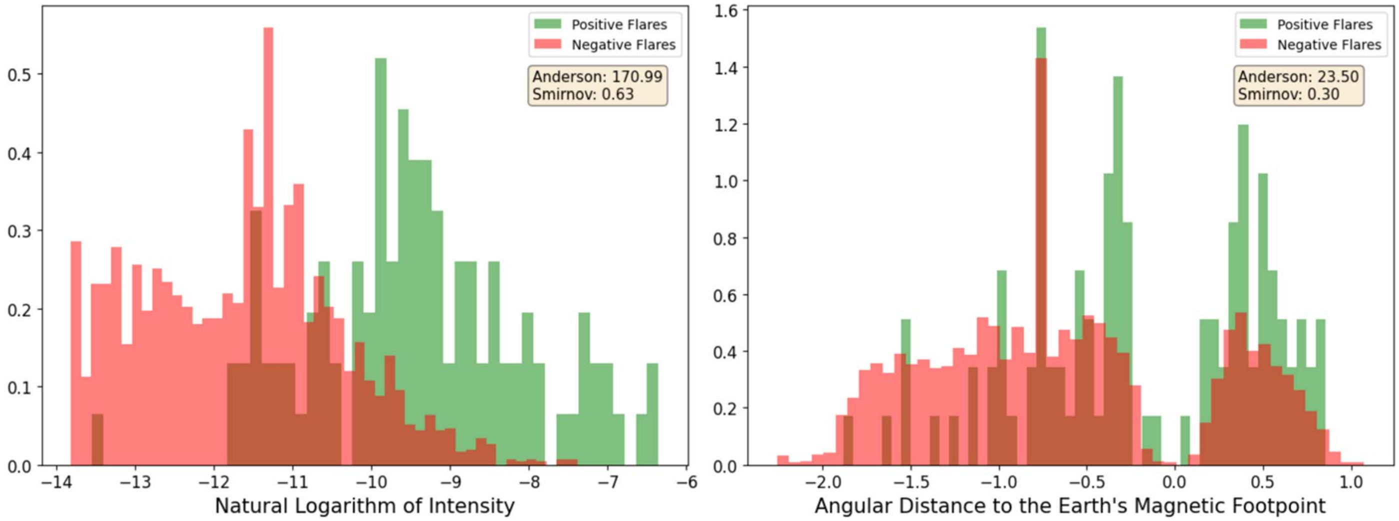

**Figure 1.** Probability density distribution of the 3356 positive (green) and 110 negative (red) events in our data set that occurred between 1996 April 4, and 2022 December 30 (SHARP/SMARP data set availability) for the logarithm of flare intensities (left panel) and angular distances to the magnetic foot point of the Earth (right panel).

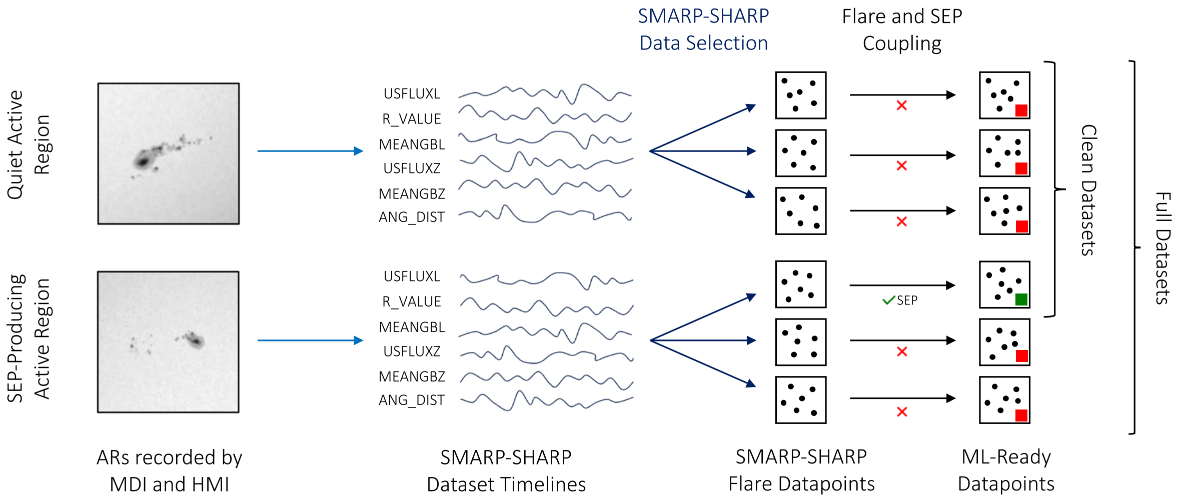

**Figure 2.** Scheme of a workflow to prepare ML-ready data sets from the SMARP and SHARP (M. G. Bobra et al. 2021; P. A. Kosovich et al. 2024). The arrows depict the previous work (light black, Section 2), the flare and SEP coupling process (black, Section 3), and the SMARP/SHARP data selection (blue, Section 4).

The procedure of preparing data sets (ML-ready) for the training and validation of the machine learning models used in this research is illustrated in Figure 3. In blue and magenta are the USFLUXL and R_VALUE timelines (Figure 3, top panel), which within a day show small fluctuations in value. Note that only two out of the six available keyword timelines (Figure 2, SMARP/SHARP data set timelines) are shown in Figure 3, annotated with the SEP starting time as defined by NOAA (red vertical line) and the starting times of the non-SEP-producing (negative) flares that occurred within AR 10180 (Figure 3, gray dashed lines). The moments of time 2, 5, and 10 hr before the start time of an SEP-producing (positive) flare are shown in green dashed lines (Figure 3, bottom only). These intervals are the different time windows this study produced results for (2, 5, 10 hr window). The forecasting window (or prediction window $t_p$, the time interval between green rectangles and lines in Figure 3) is defined as the selected time window plus the time to the first available data point. The three time windows were chosen based on the data cadence of the SOHO MDI, which is 1.6 hr (greater than the SDO HMI data cadence) and therefore 2 hr is the shortest, round, prediction window that could be used for our analysis. We set the upper limit to be 10 hr because extending the time window would lead to loss of positive events due to data availability constraints.

The matching algorithm picks the first available SHARP/SMARP data point (Figure 3, bottom panel, green rectangles) before $t_{\rm start} - x$ for $x \in [2, 5, 10]$ hr. This process is the same for all flares, regardless of whether they produced an SEP or not. These are the positive and negative data points (for three different time windows) that will comprise the ML-ready data sets discussed in this study. Every flare event has a unique prediction window $t_p \geqslant t_{\rm start} - x$. The data set has a mean forecast window of 14.21 hr for the positive flares and 12.06 hr for the negative flares. Similar forecast windows were used in previous studies (A. García-Rigo et al. 2016; A. Anastasiadis et al. 2017), allowing for reliable comparison of results.

To understand better the multidimensional problem that the ML model has to solve, we present in Figure 4 density

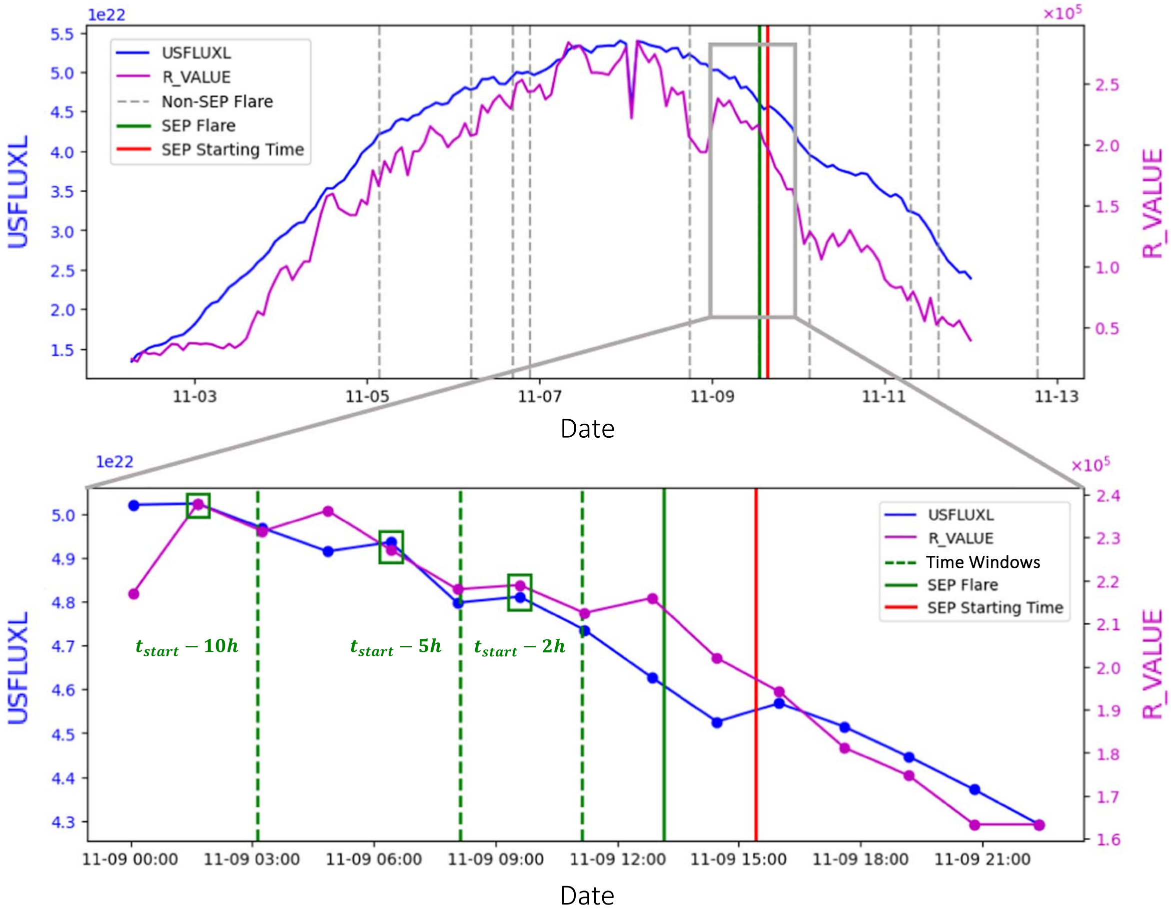

**Figure 3.** Timelines of the total line-of-sight unsigned flux (USFLUXL in Maxwells) and the unsigned flux $R$ near polarity inversion lines (R_VALUE in Maxwells) for the total tracking period of AR 10180 and the day (2002 September 11) during which an SEP occurred (bottom). Underlined (solid green) are the start time (*t_start* keyword) of the SEP producing (positive) flare of this AR and the data points that the matching algorithm selects for the ML-ready data sets. The process of selecting the data points in the green boxes is referred to as SMARP/SHARP data selection in Figure 2.

histograms of the six predictors (bold keywords, Table 1). The use of density instead of normal histogram values is essential when the populations of two different groups (positive and negative), are highly imbalanced like in our study. A strong SHARP/SMARP predictor would be one where the positive (green) and negative (red) event distributions have less overlap. The physical parameters that represent the mean value of the line-of-sight magnetic field gradient (MEANGBL; Figure 4(f)) offer less information about the eruption of an SEP. The angular distance of the AR and the magnetic foot point of the Earth, along with the unsigned flux $R$ near polarity inversion lines (R_VALUE; Figure 4(d)), provide some information about the ability to distinguish between positive and negative flares and therefore the SEPs that can reach Earth. SEPs will unlikely start early in the evolution of the AR (or west of the magnetic foot point of the earth as seen for AND_DIST_AR, Figure 4(e)). Two Gaussian distributions are visible at −0.4 and 0.4 rad from the magnetic foot point. Therefore, some predictive capability can be observed in this single parameter, which, in combination with others, could distinguish whether a flare about to erupt can lead to an SEP or not.

Similar to S. Kasapis et al. (2022), the predictors that carry most of the information beneficial to prediction tasks are those related to the unsigned magnetic flux (USFLUXL and USFLUXZ, Figures 4(a), (b)). Although for both USFLUXL and USFLUXZ the overlap between positive and negative flares is significant (still less than the rest of the predictors), these predictors can carry meaningful information for an ML method to distinguish between the two populations. The difference between their distributions is noteworthy, with USFLUXZ showing better predictive capabilities. As expected, flares produced in AR areas where there is an increased value of magnetic flux, are more likely to produce an SEP. This is even more obvious for values of unsigned magnetic flux on the line of sight that are greater than 1.75 Maxwells (USFLUXZ $\geqslant$ 1.75 Maxwells) as seen in panel b of Figure 4. A similar behavior can be observed for large values ($\geqslant$120 Gauss Mm$^{-1}$) of the line-of-sight field gradient (MEANGBZ; Figure 4(d)).

Following the SMARP/SHARP data selection processes described in this section, we have created the first group of data sets for different time windows, which include information for all flares that occurred within the tracked ARs. We proceed with creating a second group of data sets (referred to as "clean") by omitting the non-SEP-productive flares in the SEP-productive ARs as it is almost impossible for an ML method to distinguish between positive and negative data points that have such similar values. The difficulty here arises because an AR that hosts an SEP-producing flare, in most cases, also hosts multiple more non-SEP-producing flares that will have similar values with each other. Omitting the non-SEP-productive flares that erupted in SEP-productive ARs, transforms the problem, as the forecasting now concerns the AR itself (and whether it is SEP-producing or not) rather than concerning the flare

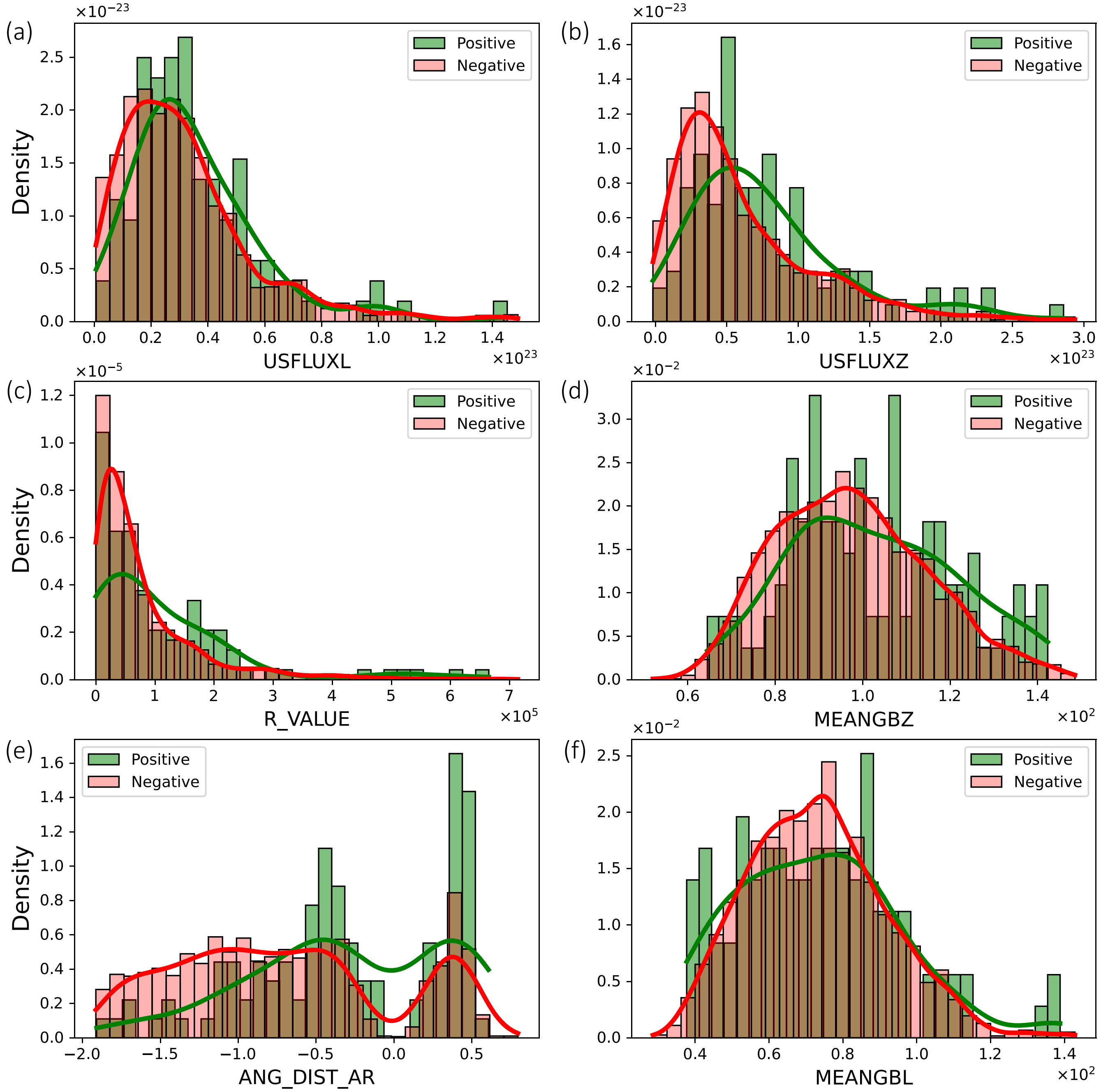

**Figure 4.** Probability density histograms for the six SHARP/SMARP predictors (Table 1) for the 10 hr window ML-ready data set for the positive (red) and negative (green) events. The red and green curves correspond to the fitted trend lines of the histogram bin values.

occurrences. In Figure 3, the data points related to the gray dashed lines ($t_{\mathrm{start}}$ of the negative flares) are the ones omitted when training our support vector machines (SVM) method on the "clean" data set. Tables 3 and 5 show how we can present our ML models with an easier problem by omitting these noise-inducing flares.

To quantify the histograms of Figure 4, the Anderson–Darling and Kolmogorov–Smirnov of the positive and negative samples in our data sets are calculated and presented in Table 3. The Anderson–Darling test (F. W. Scholz & M. A. Stephens 1987) is a statistical test that measures how well the data fits a specified distribution, with a particular focus on the tails of the distribution, making it sensitive to differences in the extremes. The Kolmogorov–Smirnov test (Massey 1951) assesses the goodness of fit between data and a theoretical distribution as a one-sample test, while the Smirnov test (V. W. Berger & Y. Zhou 2014), its two-sample counterpart used in this research, evaluates whether two data sets come from the same distribution by comparing their empirical cumulative distribution functions.

The first two columns under the "regular" category in the table represent the Anderson and Smirnov for the data set which includes all flares (not "clean"). For instance, the keyword USFLUXL demonstrates an Anderson–Darling statistic (Anderson) of 2.3164 and a Smirnov statistic of 0.1309, reflecting a moderate distinction between the two groups. In contrast, ANG_DIST_AR, with an Anderson statistic of 23.8579 and a Smirnov statistic of 0.3015, reveals a more

**Table 3**
Anderson and Smirnov Coefficients for the Six Keywords of the Regular and Clean Data Sets

| Keyword | Regular | | Clean | | Difference |
|---|---|---|---|---|---|
| | Anderson | Smirnov | Anderson | Smirnov | Anderson / Smirnov |
| USFLUXL | 2.3164 | 0.1309 | 6.8960 | 0.1813 | +4.5796/+0.0504 |
| USFLUXZ | 9.5532 | 0.2273 | 17.7214 | 0.2784 | +8.1682/+0.0511 |
| R_VALUE | 4.9641 | 0.1681 | 16.0888 | 0.2379 | +11.1247/+0.0698 |
| MEANGBZ | 3.8905 | 0.1407 | 9.4534 | 0.1885 | +5.5629/+0.0478 |
| ANG_DIST | 23.8579 | 0.3015 | 24.6131 | 0.3090 | +0.7552/+0.0075 |
| MEANGBL | 0.3643 | 0.0845 | 1.7362 | 0.1080 | +1.3719/+0.0235 |

**Note.** The last two columns include the difference in Anderson and Smirnov between the regular and clean data sets.

pronounced difference, indicating that the disparity in distributions is more significant. This pattern holds true across all data sets; keywords that exhibit higher values for both Anderson and Smirnov tests, such as USFLUXZ, R_VALUE, and ANG_DIST_AR, suggest a greater divergence between the distributions of positive and negative samples. Conversely, keywords with lower statistics, like MEANGBL_GMM, indicate a closer resemblance between the two distributions, implying a lesser degree of distinction.

The increase in both Anderson–Darling and Kolmogorov–Smirnov statistics for USFLUXL, USFLUXL, USFLUXZ, R_VALUE, and MEANGBZ, upon cleaning the data, signifies an enhancement in the ability to distinguish between the distributions of positive and negative flares. This could lead to improved discrimination power when these variables are used as features in our prediction models. However, it is important to note the very slight increase in the Anderson for ANG_DIST AND MEANGBL upon cleaning. This underlines the boundaries of predictive power within the SHARP/SMARP data set, even when the prediction problem becomes easier.

The flare intensity statistic values are indicative of the predictive power the flare information holds. The Anderson statistic for flare intensity is 170.99, substantially greater than that of ANG_DIST (23.85). Similar observations can be done for the Smirnov values proving the consistency of these results. For the physical parameter that the two data sets (flares and SHARP/SMARP) share in common, the distance to the magnetic foot point of the Earth, the Anderson of 23.50 for *ang_dist* (Table 2) is very similar to that of ANG_DIST. This proves that using the exact point the flare erupted in *ang_dist* compared to the, less accurate, AR midpoint coordinates in ANG_DIST, does not provide any additional information useful for prediction. Given the concrete knowledge these statistic tests provide, in the following Sections we will be using the SMARP/SHARP data set predictors to forecast SEP events at least half a day (14 hr) before they are observed.

## 5. Machine Learning Models and Methodology

To evaluate the predictive capabilities of the SMARP/SHARP data set for SEP prediction, we use the SVM and linear regression models available in the Scikit-Learn[7] Python library (*sklearn*). SVMs (M. A. Hearst et al. 1998; I. Steinwart & A. Christmann 2008), a supervised ML approach, are designed for both classification and regression tasks. Their primary objective is to find the optimal hyperplane that best separates different class data points in a high-dimensional space. They are particularly effective in high-dimensional spaces such as ours, as we deal with six SMARP/SHARP predictors (six dimensions) and when the classes are linearly separable. Since there is no certainty about the linear separability of our data (Figure 4), especially in their original feature space, in this research we explore different SVM kernels. The kernel trick involves mapping the data into a higher-dimensional space where it becomes linearly separable, or more separable than in the original space. The SVM kernels used in this research are: (a) the polynomial kernel (poly: finds a polynomial of the given degree to separate the data), (b) the radial basis function (RBF) or Gaussian kernel (RBF creates a landscape where data points that are close in the original space are at the peak, and those further away are down the slope) and the sigmoid kernel (sigmoid maps the similarity between data points into values between −1 and 1).

The second group of ML models is regression models (L. Fahrmeir et al. 2013), which are a set of statistical methods for estimating the relationships among variables. They are used to predict a continuous outcome variable (dependent variable) based on one or more predictors (independent variables). The most common type of regression analysis is linear (S. Weisberg 2005), where a line of best fit is determined, but other types of models include logistic (T. G. Nick & K. M. Campbell 2007), polynomial (R. M. Heiberger et al. 2009) and Ridge (D. W. Marquardt & R. D. Snee 1975) regression. While linear regression is traditionally used for predicting continuous numerical values, there are variants of linear models suitable for non-continuous or categorical data such as the ones discussed in this work (SEP/no-SEP). Logistic regression models such as the ones used in this and many other SEP works (M. Laurenza et al. 2009, 2018; L. M. Winter & K. Ledbetter 2015; A. Papaioannou et al. 2018), for instance, are designed for binary classification tasks that fit the positive/negative (SEP/No-SEP) prediction problem.

These ML models are dependent on a set of hyperparameters, which play a crucial role in determining the performance of the prediction capability. While model parameters are learned during training, hyperparameters are external to the model, they often control its overall behavior. The choice of the kernel (linear, polynomial, RBF, etc.) and their respective parameters, like the degree for a polynomial kernel or gamma for RBF, are all considered hyperparameters. Similarly, in logistic regression, the regularization strength and type (L1 or L2) are hyperparameters that can influence the model's performance. Manually selecting these hyperparameters can be suboptimal; we, therefore, employ

[7] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

**Table 4**
ACC Values for SVM and Regression Models When Trained on Six Combinations of the SMARP/SHARP Predictors for Runs That Include All Flares (Regular Data Set in Table 5), a 10 hr Time Window, and a "Balanced" Setting

| | SVM | | | | Regression | |
|---|---|---|---|---|---|---|
| SMARP/SHARP Predictors | Linear | RBF | Polynomial | Sigmoid | Logistic | Ridge |
| ALL PREDICTORS | 0.64 ± 0.10 | 0.62 ± 0.10 | 0.64 ± 0.10 | 0.62 ± 0.10 | 0.65 ± 0.09 | 0.65 ± 0.12 |
| USFLUXZ, R_VALUE, ANG_DIST | 0.66 ± 0.09 | 0.65 ± 0.10 | 0.66 ± 0.10 | 0.64 ± 0.10 | 0.65 ± 0.10 | 0.65 ± 0.09 |
| USFLUXL, MEANGB, MEANGBZ | 0.55 ± 0.10 | 0.55 ± 0.10 | 0.54 ± 0.10 | 0.53 ± 0.10 | 0.56 ± 0.10 | 0.56 ± 0.11 |
| R_VALUE, ANG_DIST | 0.67 ± 0.10 | 0.66 ± 0.10 | 0.67 ± 0.11 | 0.67 ± 0.09 | 0.66 ± 0.10 | 0.65 ± 0.10 |
| USFLUXZ, ANG_DIST | 0.66 ± 0.10 | 0.64 ± 0.11 | 0.64 ± 0.10 | 0.66 ± 0.10 | 0.66 ± 0.11 | 0.65 ± 0.10 |
| USFLUXZ, R_VALUE | 0.54 ± 0.11 | 0.58 ± 0.11 | 0.54 ± 0.08 | 0.58 ± 0.10 | 0.57 ± 0.09 | 0.55 ± 0.11 |

tools like GridSearchCV[8] from Scikit-Learn, which offers an automated and systematic approach to hyperparameter tuning. By searching through specified hyperparameter combinations and cross-validating, the GridSearchCV algorithm ensures the selection of hyperparameters that yield the best model performance. However, it is worth acknowledging the computational expense of such a brute-force approach, especially when considering extensive hyperparameter spaces or sizable data sets. In this study, we train our models using six-dimensional feature vectors in the number of low thousands; therefore, the computational expense is on the order of minutes. In the following section, SEP prediction results are produced by six ML methods: linear, RBF, polynomial, and sigmoid SVMs, along with logistic and ridge regression.

To assess the performance of each machine learning model, a suite of evaluation metrics was chosen (Table 5). Accuracy (ACC) measures the proportion of correct predictions made by the model relative to the total number of predictions. For ACC, a score of 1 indicates perfect prediction, while a score of 0.5 suggests no better than random guessing. The true skill statistic (TSS) and the Heidke skill score (HSS) both account for the skill of the model in distinguishing between the classes, factoring in both false positives and false negatives. For both TSS and HSS, a score of 1 indicates perfect skill, 0 indicates no better than random prediction, and negative values indicate inverse or contrary predictions. The false-alarm rate (FAR) measures the proportion of negative instances that were incorrectly classified as positive; a lower FAR is more desirable, with 0 being the ideal score. Lastly, the F1 Score provides a balance between precision and recall. An F1 score closer to 1 indicates better balance and performance, while a score closer to 0 suggests poor performance. These metrics provide a comprehensive evaluation framework, ensuring the model's performance is assessed from multiple vantage points.

To robustly evaluate the predictive ability of the various predictors discussed in Section 4 and to quantify model uncertainty, each model was trained and validated across 100 distinct runs. Therefore, the results reported in this work reflect the mean value and standard deviation of 100 distinct ML training and testing runs that used the same hyperparameters but different initializations. A one-to-ten (0.1) train-to-test ratio is used for the validation of every individual run. Our ML-ready data sets include 110 positive events; therefore 99 of them are used for training and 11 for testing. The methodology employed for training the aforementioned ML models was designed to account for the severe class imbalance in the data set. There are 3356 negative samples (non-SEP-producing flares) in contrast to only 110 positive samples (SEP-producing flares), resulting in an imbalance ratio of 1/30. Two types of training and validation schemes are used in our analysis: one where the positive and negative data sets are balanced (referred to as balanced) and one where the 1/30 imbalance is retained (imbalanced).

In the balanced case, every run utilizes all positive samples (110) and an equal number of negative samples, randomly chosen out of the total population (3356). Therefore the 110 chosen samples are different in each one of the 100 runs, ensuring a fair representation of the entire population. This experimental configuration (referred to as balanced in Table 5) was intentionally chosen to discern the inherent predictive capability of the predictors in a balanced, experimental scenario. In this case, testing is performed using the same amount of negative and positive samples. In the second case (labeled imbalanced in Table 5), rather than randomly picking an equal amount of negative samples, we used the entire data set (retains the 1/30 imbalance) but added a weighting factor, which during training allows for more attention to the minority class (positive). Here, the number of negative samples reserved for testing is chosen to be 30 times greater (3300 negative testing samples) than that of positive.

## 6. Results

In the pursuit to identify the most effective machine learning model for SEP prediction, a series of experiments were conducted with different combinations of the six SMARP/SHARP predictors (Figure 4). We performed tests with all possible predictor combinations. In this paper, we present the five best combinations. Across these tests, the performance of the models in terms of accuracy remained within a relatively narrow band, from 0.53 ± 0.10 for the USFLUXL, MEANGB, MEANGBZ combination (Table 4), to 0.67 ± 0.11 for the R_VALUE, ANG_DIST predictors. It is noteworthy that although the R_VALUE has lower Anderson and Smirnov values than USFLUXZ when used along with the highest Anderson predictor in Table 3 (ANG_DIST), produces slightly better results than if the ANG_DIST was combined with the USFLUXZ. This pair consistently achieved the highest accuracy across all models, reaching 0.67 ± 0.09 when training an SVM model that uses the sigmoid kernel.

Surprisingly, incorporating more predictors did not translate to enhanced performance. For instance, using all available predictors yielded accuracies such as 0.64 ± 0.10 (0.03 decrease compared to examples with the best performance, Table 4) for the linear SVM model, which is not superior to using just two predictors, as seen with R_VALUE and ANG_DIST. This can be attributed to the inclusion of

[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

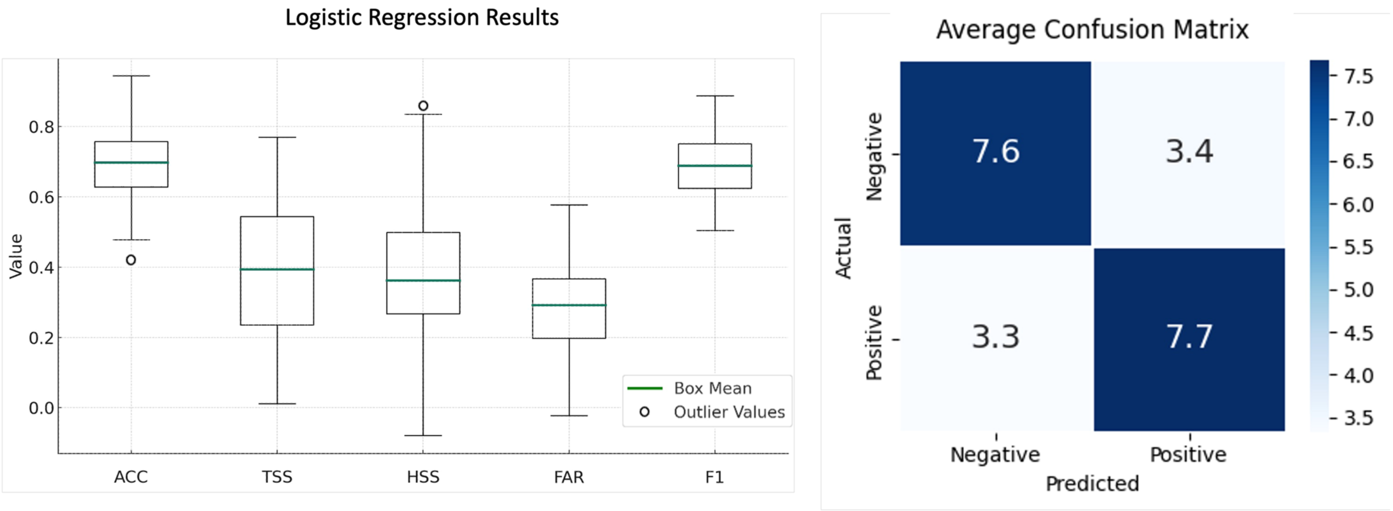

**Figure 5.** Distribution of 100 score values of ACC, TSS, HSS FAR, and F1 (left panel). Each metric is calculated using the entries of the confusion matrix produced in each run, the average over 100 runs (right panel). The values were obtained using the R_VALUE and the ANG_DIST on a logistic regression model and constitute the best SEP prediction the SMARP/SHARP data can achieve. The box range shows the interquartile range, the green line inside is the median value, the whiskers show the results range, and the circles show two outlier values. The samples are obtained using the clean data set, where non-SEP-producing flares in SEP-producing ARs are omitted and used on a balanced training method where positive and negative training and evaluation data sets are of equal size.

**Table 5**
Results for a Variety of ML Models and Training Configurations When Using the Strongest Predictors Couple, the R_VALUE and the ANG_DIST

| Setting | Training | Data Set | Model | ACC | TSS | HSS | FAR | F1 |
|---|---|---|---|---|---|---|---|---|
| Operational | Imbalance | Regular | SVM Linear | 0.56 ± 0.04 | 0.32 ± 0.10 | 0.05 ± 0.02 | 0.24 ± 0.10 | 0.71 ± 0.04 |
| S. Kasapis et al. (2022) | Imbalance | Regular | SVM Poly | 0.52 ± 0.05 | 0.01 ± 0.01 | 0.01 ± 0.01 | 0.98 ± 0.05 | 0.58 ± 0.06 |
| Semi-operational | Balanced | Regular | SVM Linear | 0.66 ± 0.10 | 0.33 ± 0.19 | 0.33 ± 0.19 | 0.26 ± 0.14 | 0.63 ± 0.12 |
| S. Kasapis et al. (2022) | Balanced | Regular | SVM RBF | 0.65 ± 0.13 | 0.33 ± 0.28 | 0.31 ± 0.26 | 0.31 ± 0.10 | 0.75 ± 0.04 |
| Semi-operational | Imbalance | Clean | SVM Linear | 0.67 ± 0.04 | 0.35 ± 0.13 | 0.08 ± 0.03 | 0.32 ± 0.12 | 0.79 ± 0.03 |
| Experimental | Balanced | Clean | SVM Linear | 0.69 ± 0.09 | 0.38 ± 0.19 | 0.38 ± 0.19 | 0.26 ± 0.14 | 0.67 ± 0.11 |
| Experimental | Balanced | Clean | Logistic Reg | 0.70 ± 0.09 | 0.39 ± 0.19 | 0.37 ± 0.19 | 0.30 ± 0.14 | 0.69 ± 0.10 |

low-quality features that may introduce noise to the ML training rather than providing valuable information. Observing the SVM model results (Table 4), it is evident that the linear, RBF, and polynomial kernels consistently demonstrate similar performance across most predictor combinations. Similarly, the regression models, both logistic and ridge, show competitive performance compared to the SVM models. Their accuracy values are in line with the top-performing SVM models, and the consistency in their performance is evident from the relatively small standard deviations.

Note that for every value tabulated in Table 4, 100 different runs are performed using an equal amount (110) of positive and negative data points (balanced), the same hyperparameters, different model initialization, and a different set of randomly picked negative data points. The training of each model is performed using 90% of the 220 (2 × 110) available data points while 10% of the data points (22) are unseen during training and reserved for evaluation (198 training vectors, 99 positive and 99 negative and 22 evaluation vectors, 11 positive and 11 negative). Similar training and testing configurations are used to produce the results of Table 5 but for the two different training processes (balanced and imbalanced) and the two different groups of data sets (regular and clean) discussed in previous sections. The training and data set selection are the factors that define whether the setting of a test is closer to simulating an operational setting (labeled as operational in Table 5) or whether the setup is more experimental (semi-operational and experimental). The experimental results are closely examined in Figure 5.

To comprehend the results in a setting that better addresses the operational needs of an SEP prediction apparatus, we adopted a different approach where each ML model is trained on the entire data set, and a weighting factor, *class_weight* (parameter in the Scikit-Learn Python library) is introduced to account for the class-imbalance problem. For the "imbalance" models (Table 5), the weight for the negative class was set to be the ratio (1/30) of positive to negative samples in the training set. This approach allows the model to emphasize the minority class during training, compensating for the lack of negative samples. The results obtained from this methodology, as presented in Table 5, reveal that the accuracy cannot remain competitive compared to the balanced and therefore experimental setting. More specifically, a significant drop in evaluation metrics such as ACC, TSS, HSS, and F1 is observed, with the largest being a 0.10 ± 0.09 decrease in accuracy. Despite this fact, the F1 score for the operational ML method imbalanced for the entire data set is higher than when we train on an experimental balanced scheme (last row, Table 5). This suggests that, in general, although the operational model's overall capability to correctly predict both positive and negative classes has decreased to almost a random pick, the model makes fewer false negatives (increased recall), and the accuracy of positive predictions (precision) remains high, even if the overall accuracy has dropped. Such trade-offs are often seen in ML binary classification.

Another two types of learning schemes are employed, one in which all available flares are included (original data set) and one where all non-SEP-producing flares that have occurred in an SEP-producing AR (Figure 2, clean data sets) have been omitted. This is done because, as demonstrated in Figure 3, it is almost impossible for an ML model trained on low-dimensional data to distinguish between a positive and a negative flare sample that has occurred almost at the same time in very close proximity. As expected, for both the balanced and imbalanced runs in Table 5, the clean data set runs exhibit higher predictive performances in the majority of the metrics analyzed. A slight increase in accuracy (+0.3 ± 0.1) is observed when training our models using a balanced data set, whereas a significant increase (+0.11 ± 0.1) can be observed when training with the clean data set a linear SVM with imbalance. This is notable because it shows that even when training with imbalance, an ML method can predict whether an AR (instead of a flare) will be SEP-productive or not.

Table 5 shows that both the operational and the semi-operational configuration results are comparable to the corresponding results by S. Kasapis et al. (2022) for similar settings. In an operational setting, an SVM model performs better (ACC increase of 0.04 ± 0.1 and F1 increase of 0.13 ± 0.1) when trained on data from two solar cycles compared to if it was trained only using the SMARP parameters. Regardless of this increase, our efforts to simulate an operational setting by conserving the inherited imbalance led to predictions that are marginally better than a random guess (Table 5, operational ACC = 0.56 ± 0.04). When shifting to a semi-operational setting, the increase in the ACC, TSS, and HSS scores for an SVM model (Table 5, third row) trained on almost double the amount of data, compared to the results of our previous work (fourth row), is within the margin of error, therefore insignificant. In this study, the FAR has decreased by 0.05 ± 0.1, showing that the increase in positive events allows the SVM models to reduce the number of false alarms. The highest accuracy achieved in S. Kasapis et al. (2022), for an experimental setting, is 0.72 ± 0.12 on a third-degree polynomial SVM when training on USFLUXL and ARDIST (equivalent of ANG_DIST in this work) results that could not be reached using the equivalent SMARP/SHARP data set predictors. In this work, the R_VALUE and ANG_DIST combination of parameters is the one that showed the best performance. This shows that the recomputed R_VALUE in the data set provided by P. A. Kosovich et al. (2024) provides meaningful information to ML models in regard to SEP production.

This work has shown us which predictors in the SHARP/SMARP data set are most useful for SEP forecasting. Another important finding of this work though, is that in realistic imbalanced problems (operational and semi-operational settings), the data set and its predictors are marginally able to predict SEPs. This is worth underlying, not only because it shows that the absence of an adequate amount of positive instances inhibits even modern solar data sets from reliably predicting SEP events, but also verifies the findings of previous works (E. Lavasa et al. 2021; M. Stumpo et al. 2021). The SMARP/SHARP data set is not suitable for SEP prediction by itself but is recommended for other tasks such as solar cycle and flare prediction, or to even complement other existing data sets used for SEP prediction applications.

The results procured from the best-performing, in terms of accuracy, run of the experimental logistic regression model in Table 5 (seventh row), trained on a balanced positive and negative setting, offer some interesting insights. The model showcases an accuracy of 0.70 ± 0.093 over 100 runs, and the TSS and HSS both register an average of 0.386 ± 0.187. The FAR of 0.303 ± 0.138 indicates a moderate rate of false alarms, while the F1 score is 0.688 ± 0.099, representing a harmonization of precision and recall. In this experimental case, where the problem is reduced to predicting whether an AR will produce an SEP, the SMARP/SHARP data set shows a notable forecasting capability, comparable to other studies (M. Núñez 2011; A. García-Rigo et al. 2016; S. Kasapis et al. 2022).

To examine the models' performance, it is essential to consider the combination of different metrics. Figure 6 shows that adjusting the threshold in logistic regression impacts the different performance metrics. By default, many ML models, including logistic regression, employ a threshold of 0.5, whereby probabilities greater than this value are designated as the positive class, and those below are designated as the negative class. Lowering the threshold from 0.65 to 0.35 generally reduces the false-alarm rate, from 0.66 ($\tau = 0.65$) to 0.118 ($\tau = 0.35$), indicating fewer false positives. However, this benefit comes with a trade-off, as accuracy drops to 0.62 at $\tau = 0.65$. The F1 Score, which represents the balance between precision and recall, is maximum when $\tau = 0.55$, suggesting this may be the best threshold for a harmonized performance. Selecting the right threshold involves balancing these trade-offs, aiming for minimal false alarms (lower FAR) or higher overall accuracy and skill (higher ACC, TSS, and HSS). The influence of the probability threshold on the metrics discussed exhibits a predictable behavior, as detailed in a series of studies (C. C. Balch 2008; M. Laurenza et al. 2009; A. Anastasiadis et al. 2017).

Lastly, a comparison of the "operational" results (Table 5, first row) to the equivalent of E. Lavasa et al. (2021) is performed, as this work has built models that have the most similarities in settings to ours. The E. Lavasa et al. (2021) "Flare_Noisy" linear SVM Model with an imbalanced data setting indicates generally better performance, as it exhibits better TSS (+0.14) and HSS (+0.36) scores. Such an increase in performance is expected as the model is trained on significantly more positive data. Comparing the FAR and the F1 scores provides further insight into the performance of the two SVM models. The current model demonstrates an FAR of 0.24 ± 0.10, significantly lower than the FAR of 0.63 ± 0.07 reported by E. Lavasa et al. (2021), suggesting that it is less prone to incorrectly predicting SEP" "operational" model achieves an F1 score of 0.71 ± 0.04, which is substantially higher than the $0.41 \pm 0.05 F1$ score of the "Flare_Noisy" linear SVM. The lower FAR indicates a more cautious prediction model that potentially reduces the number of false alerts while the higher F1 score signifies that the current model maintains a respectable balance between the ability to detect SEP events (sensitivity) and the accuracy of those detections (precision).

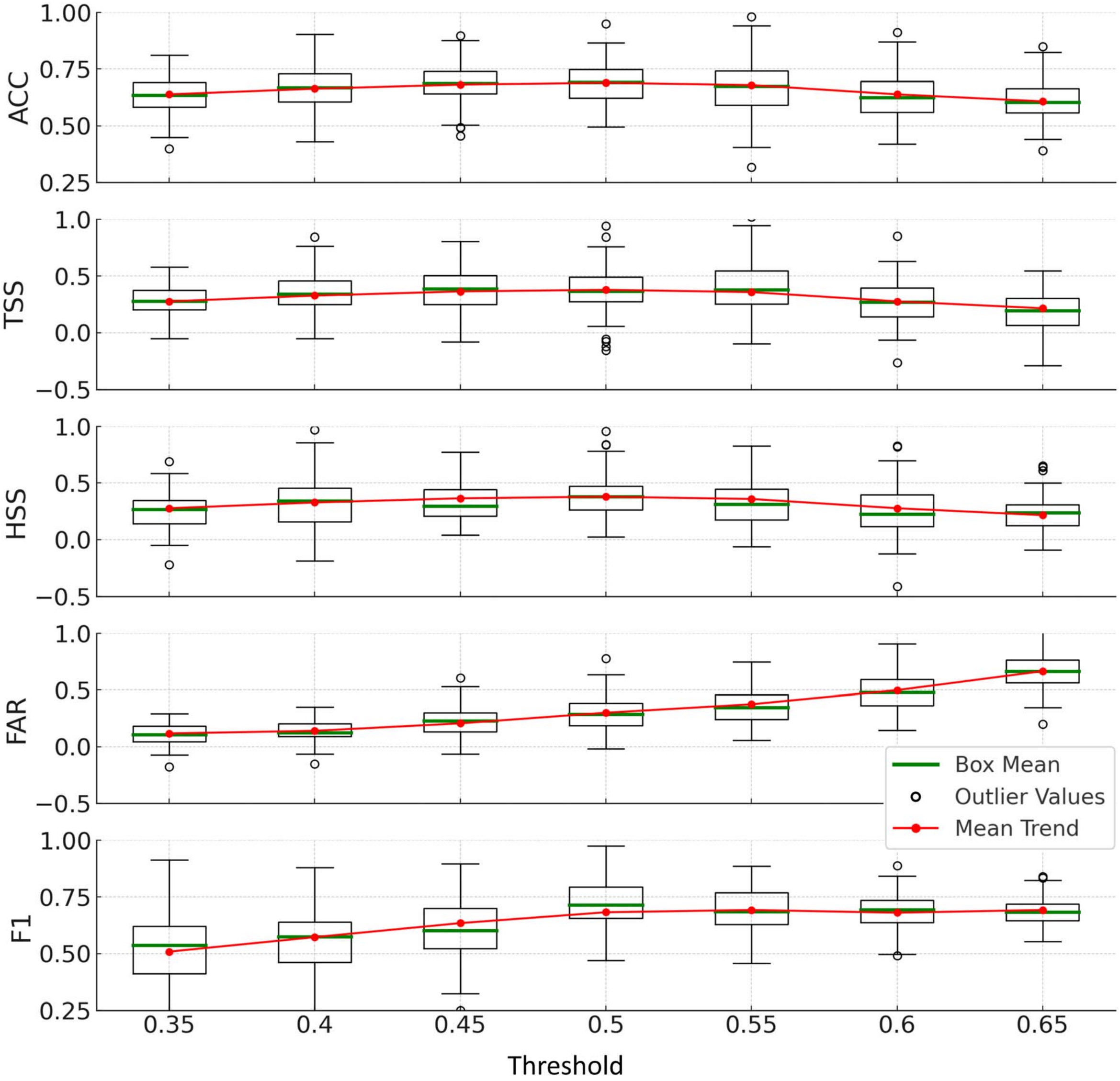

**Figure 6.** Box plots of the distribution of 100 score values of ACC, TSS, HSS FAR, and F1, for seven decision thresholds $\tau$. The values were obtained using the R_VALUE and the ANG_DIST to train an experimental logistic regression model. The box range shows the interquartile range, the green line inside it shows the median value, the whiskers show the results range, and the circles show two outlier values. These results correspond to the seventh row in Table 5 for $\tau = 0.5$.

## 7. Conclusion

To refine SEP event prediction, this work utilizes a new data set that combines the SMARP and SHARP parameters covering Solar Cycles 23 and 24 (P. A. Kosovich et al. 2024). The SMARP/SHARP data set provides physical parameters for 110 ARs that produced SEP events. Due to differences in their problem formulation, it is difficult to make fair comparisons between works on SEP prediction in space-weather literature. Despite this fact, in this work, a fair comparison is possible with S. Kasapis et al. (2022) and E. Lavasa et al. (2021) as all sets of experiments use similar setups. This work can be seen as a continuation of the work by S. Kasapis et al. (2022) with the novelty being that it: (a) extends the available data to span two solar cycles instead of one, thus increasing the rare positive instances and (b) tests the SMARP/SHARP data set on the SEP prediction while also the results interpretation reveals the new SMARP/SHARP physical values that carry SEP precursor information. The knowledge gained from this research effort allows for future work recommendations.

A set of ML models was trained on a number of data sets that differ in the number of predictors used (varying from 2 to 6), their type (USFLUX, R_VALUE, MEANGBL, USFLUXZ, MEANGBZ, and ANG_DIST), and the time window (2, 5, and 10 hr) their selected data offer. Our investigation reveals that when the SMARP/SHARP data set-based ML model faces a problem of detecting whether an AR will produce an SEP, regardless of how many flares occurred within it, the resulting accuracy of $0.70 \pm 0.09$ is comparable to the results of the author's previous study. When encountering the problem of predicting whether a flare produces an SEP or not, the predictive power of the data set diminishes. Although the results for an ML model based on the SMARP/SHARP data set appear better than if only SMARPs were used, for an operational setting, the accuracy results are only slightly better ($\mathrm{ACC} = 0.56 \pm 0.04$) than a random pick.

Our results' similarities with those in previous studies, underscores the inherent complexity of SEP prediction; even with increased data (double in volume), the ceiling of accuracy remains consistent. Interestingly, our results indicate a modest increase of 0.04 ± 0.05 in ACC but significant improvements for the TSS, HSS, FAR, and F1 metrics in the operational context, where data imbalance is introduced. The SMARP/SHARP data set contributes positively here, yet the overall accuracy achieved still denotes a considerable margin for enhancement. This suggests that the predictive capability of the data, while evident, lacks the reliability required for confident operational forecasting. The low dimensionality of the SMARP/SHARP data and the inherent imbalance of the problem hinder our models' ability to distinguish between SEP-producing and non-SEP-producing flares that occur within the same AR. This is also evident when testing different prediction windows, where the SMARP/SHARP values do not change enough between the 2, 5, and 10 hr windows, and therefore, the trained models did not produce noticeably different results. Consequently, we acknowledge the limitations of the SHARP/SMARP data set and advocate for the exploration of more sophisticated methodologies that may understand better the intricate patterns of SEP events. The authors recommend that future work should focus on: (a) using data sets that include even more positive instances (SEPs), (b) increasing the input size and the dimensions (predictors) of the data, and (c) using deeper and more sophisticated ML methods. The employment of deeper networks on continuous high-resolution SDO/HMI products has shown promising results in predicting the emergence of ARs (S. Kasapis et al. 2024); therefore similar methods should be tested on the prediction of SEP events too.

Lastly, our study shed light on the relevance of the SMARP/SHARP data set physical parameters (keywords) to SEP events. This study verifies that the total line-of-sight unsigned magnetic flux, the distance of the flare location to the magnetic foot point of the Earth, and the unsigned flux $R$ near the polarity inversion lines are physical quantities that relate to the production of SEP events. It is recommended that future studies use them for the prediction of SEP events. The importance of magnetic connectivity between the flare location on the solar disk and the Earth is also shown for an SEP to be detected.

## Acknowledgments

This work was partially supported by the NASA Heliophysics Supporting Research Grant, the NASA grants 80NSSC19K0630 and 80NSSC20K0302, and NSF grants 1639683 and 1916509. The code for this project is available in Zenodo (S. Kasapis 2024) and also at https://github.com/skasapis/SEP_Pred_SMARP-SHARP. We acknowledge the SOLSTICE Team at the University of Michigan, Department of Climate and Space Sciences and Engineering for their contributions to the methodology of this work. Lastly, we want to express our sincere gratitude to A. S. Johnson and Kathryn Whitman from the Space Radiation Analysis Group (SRAG) at the Johnson Space Center for providing us with their most updated list of SEPs.

## ORCID iDs

Spiridon Kasapis https://orcid.org/0000-0002-0972-8642
Irina N. Kitiashvili https://orcid.org/0000-0003-4144-2270
Alexander G. Kosovichev https://orcid.org/0000-0003-0364-4883
Viacheslav M. Sadykov https://orcid.org/0000-0002-4001-1295

## References

Anastasiadis, A., Papaioannou, A., Sandberg, I., et al. 2017, SoPh, 292, 21
Balch, C. C. 2008, SpWea, 6, 6
Berger, V. W., & Zhou, Y. 2014, Wiley StatsRef: Statistics Reference Online (New York: Wiley)
Bobra, M. 2017, AAS SPD meeting, 48, 111.01
Bobra, M., Hoeksema, J., Sun, X., & Team, H. M. F. 2014, SoPh, 289, 3549
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. 2021, ApJS, 256, 26
Cane, H., Richardson, I., & Von Rosenvinge, T. 2010, JGRA, 115, 8101
Chen, A., Ye, Q., & Wang, J. 2021, SoPh, 296, 150
Desai, M., & Giacalone, J. 2016, LRSP, 13, 3
Engell, A., Falconer, D., Schuh, M., Loomis, J., & Bissett, D. 2017, SpWea, 15, 1321
Fahrmeir, L., Kneib, T., Lang, S., et al. 2013, Regression Models (Berlin: Springer)
García-Rigo, A., Núñez, M., Qahwaji, R., et al. 2016, JSWSC, 6, 15
Georgoulis, M. K., Bloomfield, D. S., Piana, M., et al. 2021, JSWSC, 11, 39
Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. 1998, IEEE, 13, 18
Heiberger, R. M., Neuwirth, E., Heiberger, R. M., & Neuwirth, E. 2009, Polynomial Regression (New York: Springer), 269
Ippolito, A., Pommois, P., Zimbardo, G., & Veltri, P. 2005, A&A, 438, 705
Ji, A., Aydin, B., Georgoulis, M. K., & Angryk, R. 2020, 2020 IEEE International Conf. on Big Data (Piscataway, NJ: IEEE), 4218
Jiao, Z., Sun, H., Wang, X., et al. 2020, SpWea, 18, 02440
Kahler, S. W., & Ling, A. G. 2017, SoPh, 292, 59
Kasapis, S. 2024, skasapis/SEP_Pred_SMARP-SHARP: Algorithm for SEP prediction using SHARP-SMARP, Zenodo, doi:10.5281/ZENODO.11640610
Kasapis, S., Kitiashvili, I. N., Kosovichev, A. G., Stefan, J. T., & Apte, B. 2024, arXiv:2402.08890
Kasapis, S., Zhao, L., Chen, Y., et al. 2022, SpWea, 20, e2021SW002842
Kontogiannis, I., Georgoulis, M. K., Park, S.-H., & Guerra, J. A. 2017, SoPh, 292, 159
Kosovich, P. A., Kosovichev, A. G., Sadykov, V. M., et al. 2024, ApJ, 972, 169
Laurenza, M., Alberti, T., & Cliver, E. 2018, ApJ, 857, 107
Laurenza, M., Cliver, E., Hewitt, J., et al. 2009, SpWea, 7, 7
Lavasa, E., Giannopoulos, G., Papaioannou, A., et al. 2021, SoPh, 296, 107
Le, G.-M., & Zhang, X.-F. 2017, RAA, 17, 123
Marquardt, D. W., & Snee, R. D. 1975, Amer. Stat., 29, 3
Massey, F. J., Jr. 1951, Journ. Amer. Stat. Assoc., 46, 68
Nick, T. G., & Campbell, K. M. 2007, Topics in Biostatistics, 273 (Totowa, NJ: Humana Press)
Núñez, M. 2011, SpWea, 9, 1
O'Keefe, P. M., Sadykov, V., Kosovichev, A., et al. 2023, arXiv:2303.08092
Papaioannou, A., Sandberg, I., Anastasiadis, A., et al. 2016, JSWSC, 6, A42
Papaioannou, A., Anastasiadis, A., Kouloumvakos, A., et al. 2018, SoPh, 293, 100
Papaioannou, A., Vainio, R., Raukunen, O., et al. 2022, JSWSC, 12, 24
Robusto, C. C. 1957, Amer. Math. Month., 64, 38
Rotti, S., Aydin, B., Georgoulis, M. K., & Martens, P. C. 2022, ApJS, 262, 29
Sadykov, V., Kosovichev, A., Kitiashvili, I., et al. 2021, arXiv:2107.03911
Scholz, F. W., & Stephens, M. A. 1987, Journ. Amer. Stat. Assoc., 82, 918
Schrijver, C. J. 2007, ApJL, 655, L117
Shea, M., & Smart, D. 1990, SoPh, 127, 297
Steinwart, I., & Christmann, A. 2008, Support Vector Machines (New York: Springer)
Stumpo, M., Benella, S., Laurenza, M., et al. 2021, SpWea, 19, e2021SW002794
Wang, J., Luo, B., Liu, S., & Zhang, Y. 2023, ApJS, 269, 54
Weisberg, S. 2005, Applied Linear Regression (New York: Wiley)
Whitman, K., Egeland, R., Richardson, I. G., et al. 2022, AdSpR, 72, 5161
Winter, L. M., & Ledbetter, K. 2015, ApJ, 809, 105
Zeitlin, C., Hassler, D., Guo, J., et al. 2018, GeoRL, 45, 5845
Zhang, M., & Zhao, L. 2017, ApJ, 846, 107