Draft version June 19, 2026
For submission to the Space Science Reviews journal

Review of Machine Learning Models for Solar Energetic Particle Prediction

Spiridon Kasapis [1,2] Pouya Hosseinzadeh [3] Kathryn Whitman [4,5] Ricky Egeland [4]
Manolis Georgoulis [6,7] Angelos Vourlidas [6] Athanasios Papaioannou [8] Eleni Lavasa [8]
Anastasios Anastasiadis [8] Giorgos Giannopoulos [8] Andrés Muñoz-Jaramillo [9] Bala Poduval [10]
Irina N. Kitiashvili [2] Alexander G. Kosovichev [2,11] Viacheslav Sadykov [12]
Soukaina Filali Boubrahimi [3] Tate T. Hutchins [13,1] Hameedullah A. Farooki [1] Manuel E. Cuesta [1]
Leng Y. Khoo [1] Sungmin Pak [1] Robert Czarnota [14,1] Jamie S. Rankin [1] Jamey Szalay [1]
Mitchell M. Shen [1] Georgios Livadiotis [1] Zigong Xu [15] David J. McComas [1] Nikolaos Sarlis [16]
Dionissios Hristopulos [17] Arik Posner [4] Alec J. Engell [18] Mohammed AbuBakr Ali [19]
Ali G. A. Abdelkawy [20] Abdelrazek M. K. Shaltout,[20] M. M. Beheary,[20] Christina O. Lee [21]
Sigiava Aminalragia-Giamini [22] Constantinos Papadimitriou [22,16] Ingmar Sandberg [22] Savvas Raptis [6]
Shah Muhammad Hamdi [3] Monica Laurenza [23] Mirko Stumpo [23] Sumanth A. Rotti [24,25]
India Jackson [12] Aatiya Ali [12] Atilim Gunes Baydin [26] Nathan Schwadron [10,1]
Subhamoy Chatterjee [27] Maher A. Dayeh [27] Gelu M. Nita [11] Patrick M. O'Keefe,[28] Chun Jie Chong [28]
Paul Kosovich [11] Russell D. Marroquin [29] Berkay Aydin [30] Petrus C. Martens [24] Lulu Zhao [31]
Yang Chen [32] Yian Yu [32] Monica G. Bobra [33] Ward Manchester [31] Tamas Gombosi [31]
Ming Zhang [34] Jesse Torres [34] Philip K. Chan [34] Mohamed Nedal [35] Kamen Kozarev [36]
Peijin Zhang [37,38] Kimberly Moreland [27,39,40,41] Hazel M. Bain [42] Samuel Hart [27,43]
Michael J. Starkey [27] Alan G. Ling [44] and Simone Benella [23]

[1] *Department of Astrophysical Sciences, Princeton University, Princeton, NJ, USA*
[2] *Computational Physics Branch, NASA Ames Research Center, Moffett Field, CA, USA*
[3] *Department of Computer Science, Utah State University, Logan, UT, USA*
[4] *Space Radiation Analysis Group, NASA Johnson Space Center, Houston, TX, USA*
[5] *KBR Wyle Services, LLC, TX, USA*
[6] *Johns Hopkins Applied Physics Lab, 11100 Johns Hopkins Rd, Laurel, MD 20723, United States*
[7] *Research Center for Astronomy and Applied Mathematics of the Academy of Athens, 4 Soranou Efesiou Street, Athens 11527, Greece*
[8] *Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece*
[9] *Southwest Research Institute, Boulder, CO, USA*
[10] *Space Science Center, University of New Hampshire, Durham, NH, USA*
[11] *Department of Physics, New Jersey Institute of Technology, Newark, NJ, USA*
[12] *Physics and Astronomy Department, Georgia State University, Atlanta, GA, USA*
[13] *Department of Computer Science, Princeton University, Princeton, NJ, USA*
[14] *Department of Mathematics, Rowan University, Glassboro, NJ, USA*
[15] *Division of Physics Mathematics and Astronomy, California Institute of Technology, Pasadena, CA, USA*
[16] *Department of Physics, National and Kapodistrian University of Athens, Athens, Greece*
[17] *School of Electrical and Computer Engineering, Technical University of Crete, Chania, Greece*
[18] *NextGen Federal Systems, Morgantown, WV, USA*
[19] *National Authority for Remote Sensing and Space Science, Cairo, Egypt*
[20] *Department of Astronomy and Meteorology, Faculty of Science, Al-Azhar University, Cairo, Egypt*
[21] *Space Sciences Lab, University of California, Berkeley, CA, USA*
[22] *Space Applications and Research Consultancy, Athens, Greece*
[23] *Institute for Space Astrophysics and Planetology, Via del Fosso del Cavaliere 100, 00133, Rome, Italy*
[24] *Department of Physics and Astronomy, Georgia State University, Atlanta, GA 30303, USA*
[25] *Aryabhatta Research Institute of Observational Sciences (ARIES), Manora Peak, Nainital-263001, Uttarakhand, India*
[26] *Department of Computer Science, Oxford University, Oxford, England*
[27] *Southwest Research Institute, San Antonio, TX, USA*
[28] *Computer Science Department, New Jersey Institute of Technology, Newark, NJ, USA*
[29] *Department of Physics, University of California San Diego, La Jolla, CA 92093, USA*
[30] *Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA*
[31] *Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA*
[32] *Department of Statistics, University of Michigan, Ann Arbor, MI, USA*

[33] Office of Data and Innovation, State of California, Sacramento, CA
[34] Department of Electrical Engineering and Computer Science, Florida Institute of Technology, Melbourne, FL, USA
[35] Astronomy and Astrophysics Section, School of Cosmic Physics, Dublin Institute for Advanced Studies, DIAS Dunsink Observatory, Dublin D15 XR2R, Ireland
[36] Institute of Astronomy of the Bulgarian Academy of Sciences, Sofia, Bulgaria
[37] Center for Solar-Terrestrial Research, New Jersey Institute of Technology, Newark, NJ 07102, USA
[38] Cooperative Programs for the Advancement of Earth System Science, University Corporation for Atmospheric Research, Boulder, CO, USA
[39] CIRES, University of Colorado Boulder, Boulder, CO, USA
[40] Space Weather Prediction Center, NOAA, Boulder, CO, USA
[41] Department of Physics and Astronomy, College of Science, The University of Texas at San Antonio, San Antonio, TX, USA
[42] Space Weather Prediction Center, National Oceanic and Atmospheric Administration, Boulder, CO, USA
[43] The University of Texas at San Antonio, San Antonio, TX, USA
[44] Atmospheric and Environmental Research, Inc., MA, USA

ABSTRACT

Solar energetic particle (SEP) events have attracted increasing attention due to their significant radiation hazards for aviation, spacecraft electronics, and human missions beyond Earth's magnetosphere. From a scientific perspective, SEP events are intriguing because they arise from a set of physical processes extending from the solar surface and corona through the heliosphere, offering insight into particle acceleration and transport mechanisms that are widely applicable across astrophysics. Therefore, advancing our ability to understand and predict SEP events is essential both for deepening our knowledge of such mechanisms and for safeguarding space technologies and exploration. Traditionally, researchers have modeled SEPs using physics-based simulations and empirical methods. More recently, machine learning (ML) has emerged as a new tool for understanding and predicting SEP events. The purpose of this manuscript is to review the currently available ML models for SEP prediction, identify the datasets used for training, compare their architectures, inputs, and outputs, and, based on these insights, outline good practices and recommendations for future research.

1. INTRODUCTION

The heliosphere, the region dominated by the Sun's influence, can be seen as a complex system of interconnected subsystems (Engelbrecht et al. 2022; Cohen et al. 2026). Within this domain, the constantly changing magnetic activity of the Sun shapes what is known as space weather (e.g., Temmer 2021; Gopalswamy 2022) in combination with the galaxy's hazards via galactic cosmic rays (Rankin et al. 2022). Space weather disturbances, particularly their more hazardous aspects, are driven by highly energetic solar events such as flares, coronal mass ejections (CMEs), and solar energetic particle (SEP) events (e.g., Papaioannou et al. 2016; Buzulukova & Tsurutani 2022). These high-energy phenomena are critical contributors to technological disruptions in space and on Earth, underscoring the importance of understanding and predicting them to mitigate their potentially severe impacts (Georgoulis et al. 2024).

SEP events are characterized by the rapid acceleration and release of high-energy electrons (e.g., Mitchell et al. 2025), protons and heavier ions (McComas et al. 2019; Cohen et al. 2021a,b; Pak et al. 2025) into the heliosphere (Desai & Giacalone 2016; Reames 2021). These particles, accelerated by transient solar phenomena such as solar flares and CMEs, exhibit energies spanning from keV to GeV (Reames 2013). Once energized, SEPs propagate along interplanetary magnetic field lines, creating complex spatial and temporal distributions influenced by the physical characteristics of their sources, e.g., diffusive shock acceleration (Axford et al. 1977; Bell 1978; Blandford & Ostriker 1978; Drury 1983), and by their transport through the interplanetary medium (Zank et al. 2015; Chhiber et al. 2021; Subashchandar et al. 2025; Cuesta et al. 2025).

SEP events pose significant challenges to space weather forecasting due to their rapid onset and variability and potential impact on both technological systems and human health (e.g., Tobiska et al. 2015; Mishev et al. 2015; Miroshnichenko 2018). SEPs are central to space weather concerns as they represent a major radiation hazard for astronauts, particularly during extravehicular activities, and for passengers and crew on high-latitude flights (Cucinotta et al. 2013; Tobiska et al. 2015). These high-energy particles can also damage satellite electronics, disrupt communication systems, and impair navigation and power infrastructure on Earth (Schrijver et al. 2015). For interplanetary missions, SEP events present critical risks, especially for astronauts on the Moon or Mars, where the lack of a planetary

magnetic field and thin or non-existent atmospheres offer limited shielding from solar radiation (Zeitlin et al. 2013). As the National Aeronautics and Space Administration (NASA) and other space agencies plan long-term missions to the Moon and Mars, accurate forecasting of SEP events has become imperative for ensuring mission success and the safety of human explorers (Neukart 2024; Creech et al. 2022).

Despite decades of research, forecasting SEP events with high confidence remains an open challenge. Traditional forecasting approaches span from physics-based models —which simulate particle acceleration and transport processes, but often require substantial computational resources— to empirical and statistical models that provide faster predictions, but rely heavily on historical correlations (Whitman et al. 2023). Recently, machine learning (ML) has emerged as a powerful alternative tool, capable of uncovering nonlinear patterns across diverse solar and heliospheric datasets (Aminalragia-Giamini et al. 2021; Neukart 2024; Kasapis et al. 2025b). ML has created a new area of heliophysics research (Nita et al. 2020; Berger et al. 2021) and a new community of researchers (Camporeale & of ML-Helio 2020; Narock et al. 2022). This growing body of work includes a recent review of empirical and physics-based models that also highlights the expanding role of ML techniques in space-weather and specifically in SEP forecasting (Papaioannou et al. 2025). ML methods hold promise for improving both the accuracy and the speed of SEP forecasts, particularly as the volume of space-based observations continues to grow.

The application of ML to the prediction of SEPs faces several challenges, including the rarity of SEP events (typically an SEP event is defined as $\geq$ 10 MeV particles surpassing a 10 pfu limit[1]) and in particular the high-energy ones (Waterfall et al. 2023), the severe class imbalance between SEP and non-SEP cases (which becomes more pronounced as the particle's energy increase), and the need to integrate heterogeneous and often sparse data sources such as proton fluxes, solar images, flare catalogs, and active region (AR) properties (Chatterjee et al. 2024; Hosseinzadeh et al. 2025; Papaioannou et al. 2025). In addition, because most ML-based SEP prediction efforts remain proof-of-concept, researchers have adopted widely varying validation metrics, input data choices, target definitions, and modeling setups, which makes meaningful comparison between studies difficult. Overcoming these limitations requires careful and more coordinated model design, innovative data augmentation strategies, and the incorporation of domain knowledge to ensure both predictive skill and physical interpretability (Lavasa et al. 2021; Sadykov et al. 2021; Ali et al. 2024; Hosseinzadeh et al. 2024a).

In this paper, we provide a comprehensive review of ML models developed for SEP prediction. Twenty four approaches are categorized based on their model architectures, input data, and output predictions (Section 2 and Appendix A). The datasets commonly used to train and validate these ML models (Section 3 and Appendix B) are identified and their performance and limitations are discussed (Section 4). An effort is made to systematically compare existing models, assess the current state of the field, identify open challenges, and outline opportunities for advancing SEP forecasting to operational use. Moving beyond cataloging existing SEP forecasting approaches (as already done by Whitman et al. 2023), the current review is focused on ML applications and synthesizes common trends, limitations, and emerging directions identifying both current capabilities and remaining challenges that need to be addressed before SEP ML forecasting becomes completely operational. By clarifying or capturing the state of the field and highlighting opportunities for future progress, this work aims to lead the development of next-generation SEP prediction systems that combine predictive skill, physical interpretability and operational reliability. Ultimately, this work provides a cartography of the current state of SEP-prediction research using ML and to guide future efforts toward more reliable, interpretable, and operationally useful approaches for space-weather forecasting, as discussed in Section 6.

2. OVERVIEW AND CATEGORIZATION OF ML MODELS

To facilitate a systematic comparison of ML models for SEP prediction, we organize the descriptions of 24 models (Table 1) identified in the English literature into three major categories: *Architecture*, *Input*, and *Output* (Table 3). Each category contains several subfields that capture aspects of the model's design. The *Architecture* section includes descriptors such as algorithm type and complexity; *Input* covers data shape, physical type, historical depth, diversity, class imbalance, and sample characteristics; and *Output* describes the prediction format, triggering mechanism, and forecast window. This type of categorization is followed throughout Appendix A, where single-page model descriptions

[1] https://www.swpc.noaa.gov/products/goes-proton-flux

Section	Model	References	Access Links	Type	Complexity
A.1	XGBoost	Ali et al. (2024)	SEP List and Data	Gradient Boosting	2
A.2	STSF	Rotti et al. (2024a), Rotti et al. (2024b)	Model and Data	Time Series Classifier	4
A.3	SMARP-SHARP	Kasapis et al. (2022), Kasapis et al. (2024)	Model, SEP List and Data	Linear SVM	7
A.4	AA	Lavasa et al. (2021)	Model and Data	Random Forest	8
A.5	ESPERTA	Laurenza et al. (2009, 2018, 2024), Alberti et al. (2017, 2019), Stumpo et al. (2021), Benella et al. (2023)	N/A	Logistic Regression	12
A.6	UMASEP	Núñez (2011)	CCMC SEP Scoreboard	Regression Tree Ensemble	20
A.7	UMASOD	Núñez & Paul-Pena (2020)	N/A	Decision Tree	30
A.8	MS-SEP	Ali et al. (2025)	Model and Data 1, Data 2, Data 3	Random Forest	52
A.9	CART	Boubrahimi et al. (2017)	N/A	Decision Tree (CART)	61
A.10	RH	O'Keefe et al. (2024)	SPE Catalog	Random Ensemble of NNs	202
A.11	SSEP	Jackson & Martens (2024a), Jackson & Martens (2024b)	Model and Data	Random Survival Forests	300
A.12	SEP-C	Torres et al. (2022)	Model and Data 1, Data 2	Neural Network	780
A.13	CANN	Sadykov et al. (2021)	SPE Catalog and Model	Custom Architecture NN	1,243
A.14	SEP-E	Torres et al. (2025)	Model and Data 1, Data 2	Neural Network	1,530
A.15	SPRINTS	Engell et al. (2017)	Outputs	MLP	5,401
A.16	TSF	Hosseinzadeh et al. (2024a)	Model	Time Series Forest	15,000
A.17	UDM	Hosseinzadeh et al. (2024b)	Model	Time Series Forest	15,548
A.18	UNSPELL	Aminalragia-Giamini et al. (2021)	Data 1, Data 2	NN Ensemble	81,120
A.19	TS-HOG-TB	Hosseinzadeh et al. (2025)	Model	Ensemble Method	100,000
A.20	SEPNET	Yu et al. (2025)	Model	Transformer	130,000
A.21	BiLSTM-SEP	Nedal et al. (2023)	Model and Data 1, Data 2, Data 3	BiLSTM NN	333,699
A.22	MEMPSEP	Chatterjee et al. (2024), Dayeh et al. (2024), Moreland et al. (2024)	Model and Data	Convolutional NN	6,092,617
A.23	PSPSP	Hutchins & Kasapis (2026)	Code and Data	Neural Network	13,814,081
A.24	EPREM-S	Baydin et al. (2023)	Model and Data	Feed-Forward NN	285,881,344

Table 1: List of ML models that predict SEP events. The first two columns reference the Appendix A subsection of the models' description and their name. The following columns include references to the models, links to the associated code and data (if available), and the models' *Type* and *Complexity* descriptors. Models are summarized in this table —and in the manuscript as a whole— in order of complexity (number of trainable parameters), beginning from the least complex models. A list of acronyms (including the acronym names of the models) along with their definitions is available in Appendix D.

are provided, based on the submissions offered by the authors of each model. The quantitative and qualitative values for each model based on such a categorization are prescribed in Tables 6-29. These values were again submitted by the modelers through responding to the form presented in Appendix C.

The *Architecture* subgroup captures the structural and computational characteristics of each SEP prediction system. It includes the *Model Type*, a categorical descriptor indicating the algorithm class —such as Support Vector Machines (SVMs), random forests, deep Neural Networks (NNs), or Long Short-Term Memory (LSTM) methods— which defines the model's learning strategy and architecture. Complementing this is the *Model Complexity*, a numerical value representing the total number of trainable parameters, i.e., the number of internal weights adjusted during training (often referred to as "model weights"). This scalar metric reflects the model's depth and capacity, offering insight into its expressiveness, computational cost, and potential for overfitting. In practice, this index spans several orders of magnitude: shallow models such as SVMs or random forests typically involve only tens to a few hundred trainable parameters, whereas modern deep NNs can contain tens of millions of parameters.

The *Input* subgroup captures the structure and physical nature of the data used to train the SEP prediction models. The *Input Shape* is a categorical descriptor that defines the dimensionality of each individual sample—whether it is point-like (0D), time series (1D), spectra (1D), or imagery (2D). This classification reflects the format of a single event, not the overall dataset structure. The *Input Type* is also categorical and refers to the physical quantity represented in the input, such as magnetic fields, extreme ultraviolet (EUV) or X-ray imagery, electric fields, white-light observations, radio measurements, and others. Beyond structure and physical type, several quantitative subcategories describe the statistical and temporal properties of the input dataset. *Input History* refers to the total time span (in years) covered by the training data, *Input Diversity* captures the total number of positive (SEP-producing) and negative events (non SEP-producing) used for training, while *Input Imbalance* quantifies the rarity of SEP-producing events as a fraction (0 to 1) of the total sample set. Additionally, *Input Sample Size* measures the data volume (in bytes) of a single event, and *Input Sample Coverage* indicates the temporal duration (in hours) represented by each input sample. These metrics are useful for understanding the dataset richness, bias, and the temporal resolution of each model's learning process.

Lastly, the *Output* subgroup describes the nature and format of the predictions produced by each SEP model. The *Output Prediction* is a categorical descriptor that defines the kind of prediction made, such as binary classification (SEP vs. non-SEP, all-clear warnings), regression of time-series quantities (onset or peak time forecasts) and probability estimates. Models may fall into multiple categories depending on how their outputs are processed or interpreted. The *Output Type* qualitative category (Triggered vs. Continuous) further distinguishes models based on their operational logic: triggered models (often referred to in literature as post-eruptive) issue predictions in response to specific solar events (e.g., flares or CMEs), while continuous models provide ongoing forecasts based on ambient solar or heliospheric conditions (often referred to as pre-eruptive). Finally, *Forecast Window*, a quantitative metric, defines the forecast window, the time span over which the prediction is valid or expected to occur, such as SEP onset within the next 7 hours or an all-clear prediction for the next 24 hours.

The qualitative descriptors of each corresponding model that exists in the current literature are prescribed in Table 3 and Figure 3 of Section 4. The quantitative descriptors are used to create Figures 2 and 4 in an attempt to compare the different models with each other and map the state of current research in the field. In Appendix A, each model and its descriptors are summarized in more detail.

3. DATASETS FOR SEP PREDICTION USING ML

ML applications for SEP prediction rely fundamentally on the quality, relevance, and structure of the datasets used for training and evaluation. Unlike traditional modeling approaches, ML methods require large volumes of labeled data that capture the complexity of SEP-related phenomena, including solar activity indicators, particle flux measurements, and contextual heliospheric conditions. The choice of dataset directly influences model performance, generalizability, and interpretability. Subsections B.1-B.5 of Appendix B summarize the key datasets that have been developed and used for SEP prediction using ML, highlighting their characteristics, input features, target variables, and typical use cases. A list of these five datasets, along with the links to access them, can be found in Table 2. Note that all models presented in this manuscript are supervised, therefore they require datasets that provide event labeling. Future studies could utilize these datasets, without their labels, to explore the capabilities of unsupervised ML approaches for SEP prediction.

It should be noted that this list is not exhaustive, but rather a list of datasets curated for the specific task of SEP prediction. A majority of the works in Table 1 and Appendix A have used data from various sources for which although extensive processing might have taken place, they have not been published as peer reviewed publications. To avoid limiting the datasets list to peer reviewed publications, Table 1 contains (if available from the modelers) links to their datasets, repositories and code. In summary, the 24 studies have used data from nine different spacecraft as

Table 2: List of datasets that were created with the purpose of being used for SEP prediction. The title of the relevant publication, the associated Digital Object Identifier (DOI), the respective Appendix B dataset description and useful links to access the datasets (if available) are included.

Developer & Title	Publication Title	Description and DOI	Dataset Access
Kimberly Moreland (MEMPSEP-III Dataset)	MEMPSEP-III. A Machine Learning-Oriented Multivariate Data Set for Forecasting the Occurrence and Properties of Solar Energetic Particle Events Using a Multivariate Ensemble Approach	Appendix B.1 10.1029/2023SW003765	Zenodo
Pouya Hosseinzadeh (MTS-SEP Dataset)	Improving Solar Energetic Particle Event Prediction through Multivariate Time Series Data Augmentation	Appendix B.2 10.3847/1538-4365/ad1de0	GitHub
Sumanth A. Rotti (GSEP Dataset)	Integrated Geostationary Solar Energetic Particle Events Catalog: GSEP	Appendix B.3 10.3847/1538-4365/ac87ac 10.3847/1538-4365/acdace	Harvard Dataverse
Paul Kosovich (SHARP-SMARP Dataset)	Time series of magnetic field parameters of merged MDI and HMI space-weather active region patches as potential tool for solar flare forecasting	Appendix B.4 /10.3847/1538-4357/ad60c3/	Google Drive
Kathryn Whitman (CLEAR Dataset)	CLEAR SEP Benchmark Dataset	Appendix B.5 CLEAR Benchmark Website	Description and Dataset

seen in Figure 1, with an overwhelming majority (21/24) using the National Oceanic and Atmospheric Administration (NOAA) Geostationary Operational Environmental Satellite (GOES) series (Rodriguez et al. 2010; Hu & Semones 2022; Sellers & Hanser 1996), many times complemented with data from other satellites such as the Solar and Heliospheric Observatory (SOHO; Domingo et al. 1995b,a), the Solar Dynamics Observatory (SDO; Pesnell et al. 2012), the Wind (Von Rosenvinge et al. 1995) spacecraft, the Advanced Composition Explorer (ACE; Stone et al. 1998), the Deep Space Climate Observatory (DSCOVR; Burt & Smith 2012) and the Solar TErrestrial RElations Observatory (STEREO; Kaiser et al. 2008). The BiLSTM-SEP and PSPSP models have also used data from the Interplanetary Monitoring Platform (IMP; Simunac & Armstrong 2004) and the Parker Solar Probe (PSP; Fox et al. 2016; Raouafi et al. 2023). Some studies such as MEMPSEP, SEPNET and ESPERTA have also used ground-based datasets and relevant event catalogs such as the Space Weather Database Of Notifications, Knowledge, Information (DONKI) developed at the Community Coordinated Modeling Center (CCMC[2]) and the Low-Frequency Array (LOFAR; van Haarlem et al. 2013).

The majority of missions from which data are used for ML-aided SEP predictions reside either in Earth orbits —such as GOES in geostationary orbit and SDO in inclined geosynchronous orbit— or at the Sun–Earth Lagrange 1 (L1) point, where SOHO, WIND, ACE, and DSCOVR operate. These spacecraft therefore primarily sample the space environment relevant to geoeffective events. A geoeffective SEP event generates particles that reach Earth with sufficient intensity and energy to produce measurable impacts on the near-Earth space environment. Only a handful of studies (PSPSEP and MEMPSEP) use data from spacecraft located away from the Earth–L1 system, such as the PSP, which operates in a highly eccentric heliocentric orbit deep in the inner heliosphere and STEREO, which is in a heliocentric orbit near Earth's orbit. These are also the only studies that have used imagery as inputs to their models (Table 3, 2D column). Since most satellites in Earth orbit or at L1 downlink their measurements in near-real-time (e.g., SDO, SOHO, WIND) and some even in real-time (GOES, ACE, DSCOVR), the majority of existing SEP-prediction studies could in principle evolve toward operational forecasting using near-real-time inputs. What constitutes an operational model, and the criteria for determining operational feasibility, are discussed in Section 5. Despite the availability of multiple SEP and solar datasets, inconsistencies in event definitions, preprocessing pipelines, and temporal coverage —stemming from the differences in the modelers' dataset selection— remain significant barriers to fair model comparison, as discussed later in this manuscript. Continued efforts toward dataset harmonization and shared validation resources will be essential for advancing ML-based SEP forecasting.

[2] https://ccmc.gsfc.nasa.gov/

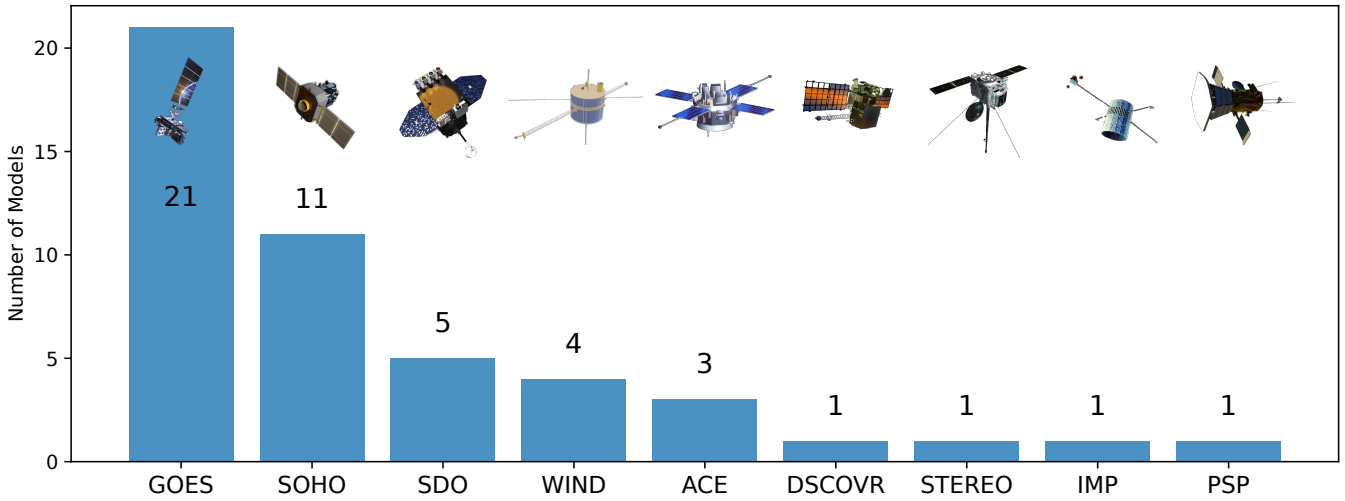


Figure 1: Histogram illustrating the number of SEP forecasting models (from 24 total) that utilized space-based observations from each respective spacecraft.

4. DISCUSSION

The past decade, the heliophysics community has explored a broad set of ML models, which have different architectures, use different data, and are at varying levels of maturity. Predicting SEP events using ML is an emerging field, and most studies are proofs of concept, compared to other physics-based and empirical models that are well-established and already provide real-time operational forecasts (Whitman et al. 2023). Table 3 provides a community-driven overview of existing SEP ML models, summarizing their architectures, inputs, and outputs based on the grouping described in Section 2 and the detailed questionnaire completed by model developers and presented in Appendix C. While this study makes every effort to compile a comprehensive list of models developed to date, it reflects only the English-language literature and the state of the field prior to 2026. As with any rapidly evolving research area, new ML-based SEP prediction approaches continued to appear as this work was being written. Consequently, the conclusions drawn here should be interpreted as a snapshot of an active and continually developing domain rather than a definitive or static inventory.

Our review compares three key aspects of these works: the inputs, the ML models' architectures trained on those inputs, and their outputs. SEP prediction approaches employ a wide range of model architectures, often driven by available input data and the desired output. Our first goal is to map the complexity of the different ML model architectures (Architecture column of Table 3 and Section 4.1). Furthermore, by comparing the inputs (Input column of Figure 3 and Section 4.2), we can understand which space- or ground-based observations the community has used, which physical quantities have proven most useful, which datasets are more accessible or easier to work with, and whether there is valuable information that has been overlooked for SEP forecasting tasks. However, comparing the models' outputs (Output column of Table 3 and Section 4.3) highlights the current status of the field: most studies explore different forecasting setups that produce different outputs. As a result, there is no standardized prediction output across models, making it difficult to compare results directly, since different studies aim to predict fundamentally different quantities. Some works predict the expected SEP onset time at Earth (e.g., SSEP, MEMPSEP), others provide all-clear predictions (e.g., BiLSTM-SEP, CANN), and several estimate the probability that a flare will produce an SEP (e.g., SHARP-SMARP, TSF, UDM, AA). Certain approaches produce entirely different types of outputs, such as inner-heliosphere particle intensity profiles (e.g., PSPSP) or surrogate physics-based simulations (e.g., EPREM-S). Additionally, some studies perform pre-eruptive forecasting while others rely on post-eruptive, triggered prediction setups. These diverse output targets make direct comparison between models inherently challenging or even impossible. Sections 4.1-5 examine the current landscape by comparing model architectures, inputs and outputs and then outline paths towards improved comparability and operational readiness.

4.1. *Model Architectures*

Table 3: Qualitative taxonomy of the SEP ML models. This table presents a summary of the qualitative values listed in Tables 6-29, for the three model aspects discussed in Section 4: a) *Architecture* (Section 4.1), b) *Inputs* (Section 4.2) and *Outputs* (Section 4.3).

Architecture						Inputs													Output				
Model Name	Type					Shape			Type										Prediction			Type	
	Neural Network	Forest/Decision Tree	Linear/Logistic Regression	Support Vector Machine	Ensemble	0D	1D	2D	Soft X-ray	Proton Flux	Solar Wind	Space-Based Radio	Ground-Based Radio	Electron Flux	Flare Location	Magnetic Fields	Coronagraphs	EUV Imagery	Classification	Probability	Regression (Time-Series)	Triggered	Continuous
A.1 XGBoost		✗			✗		✗		✗	✗									✗				✗
A.2 STSF		✗			✗		✗		✗	✗									✗				✗
A.3 SMARP-SHARP			✗	✗		✗										✗			✗	✗		✗	
A.4 AA		✗				✗			✗								✗		✗			✗	
A.5 ESPERTA			✗			✗			✗			✗	✗		✗				✗	✗		✗	
A.6 UMASEP		✗	✗		✗		✗		✗	✗									✗		✗		✗
A.7 UMASOD		✗					✗		✗			✗							✗			✗	
A.8 MS-SEP		✗					✗		✗			✗					✗		✗			✗	
A.9 CART		✗					✗		✗	✗									✗			✗	
A.10 RH	✗				✗	✗			✗						✗				✗			✗	
A.11 SSEP		✗				✗									✗					✗		✗	
A.12 SEP-C	✗					✗				✗	✗						✗		✗	✗	✗		✗
A.13 CANN	✗					✗			✗	✗			✗			✗			✗	✗			✗
A.14 SEP-E	✗						✗			✗				✗					✗	✗	✗		✗
A.15 SPRINTS	✗					✗			✗						✗				✗	✗		✗	
A.16 TSF		✗				✗	✗			✗									✗			✗	
A.17 UDM		✗					✗			✗								✗	✗			✗	
A.18 UNSPELL	✗				✗	✗			✗						✗				✗	✗		✗	
A.19 TS-HOG-TB		✗		✗	✗		✗			✗								✗	✗			✗	
A.20 SEPNET	✗						✗		✗							✗			✗	✗			✗
A.21 BiLSTM-SEP	✗						✗		✗	✗	✗		✗			✗			✗	✗	✗		✗
A.22 MEMPSEP	✗				✗	✗	✗	✗	✗		✗	✗		✗		✗	✗	✗	✗	✗	✗	✗	
A.23 PSPSP	✗						✗	✗		✗	✗							✗	✗		✗		✗
A.24 EPREM-S	✗						✗			✗													✗

Table 3 presents a summary of the qualitative responses submitted by SEP modelers with regards to the model's architecture, input and outputs. The questionnaire researchers responded to is available in Appendix C. The values checked within Table 3 for each model are also those presented in Tables 6-29. The left-most group of columns in Table 3 (Architecture) presents a summary of the model types the community has used.

Many models (11/24) use some type of forest or decision tree architecture. The type ranges from a gradient-boosted decision tree architecture (i.e., XGBoost) to the Random Forest, a bagging ensemble classifier built on decision trees (AA) and to the Classification and Regression Tree (CART) model, which are simple decision trees that use splitting criteria to identify feature thresholds that best separate SEP and non-SEP events. Such methods are appealing because they perform well with relatively small datasets, are robust to noise and imbalance, and offer easier interpretability through feature importance analysis. Therefore, these remain competitive baseline approaches for SEP

prediction. Similarly, NNs are also a popular type of model (11/24). A wide variety of NN architectures have been used. For example, BiLSTM-SEP uses a Bidirectional LSTM network, a type of Recurrent Neural Network (RNN), the MEMPSEP model uses Convolutional Neural Networks (CNNs), SEPNET uses Transformers and approaches such as UNSPELL and RH use ensembles of NNs. Seven of the models (7/24) use some type of ensemble architecture, a method where multiple models are combined to produce a single, usually better, prediction than any individual model could achieve on its own. Fewer models (5/24) have used lower-complexity architectures, such as regression models or SVMs. Nonetheless, they remain valuable and can act as useful baselines or exploratory tools for identifying useful information about physical parameters.

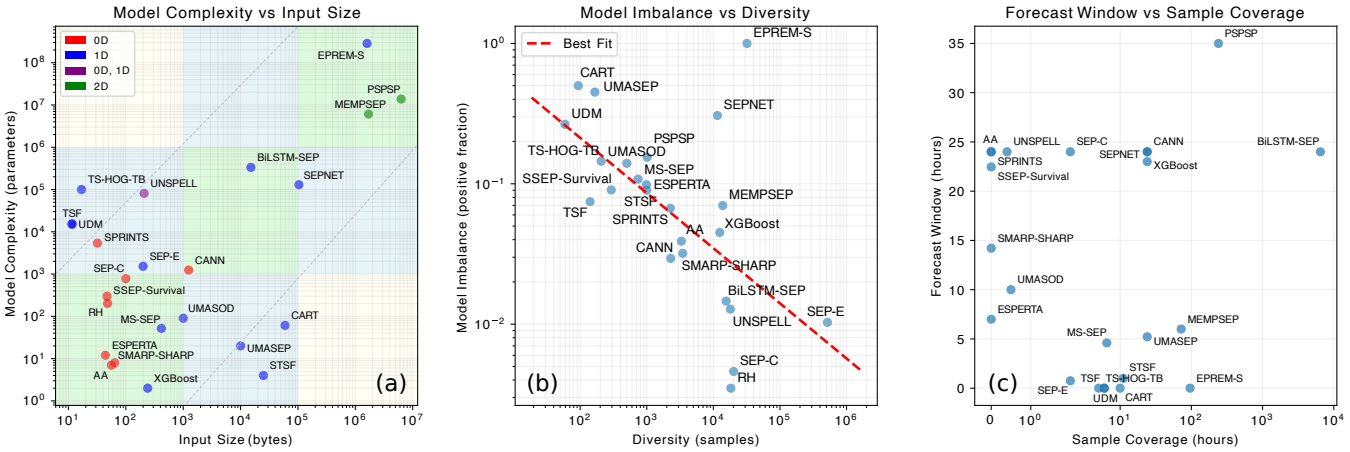


Figure 2: Plots of the quantitative categorization submissions of the SEP modelers tabulated in Tables 6-29. Panel (a) presents the model complexity against the input size in bytes. Panel (b) presents the input's imbalance against the input diversity. Panel (c) presents the forecast window against the sample coverage.

The models examined in this review differ not only in architecture but also in complexity. Here, *Model Complexity* refers to the number of free parameters (trainable weights) contained within each model. Figure 2a illustrates this complexity as a function of *Input Size*. For visualization, models are grouped into three categories: low complexity (1–1,000 parameters; bottom row), medium complexity (1,000–100,000 parameters; middle row), and high complexity ($\geq 100,000$ parameters; top row). A clear observation from this distribution is that, because ML-based SEP prediction is still a nascent field, most models (12/24) fall into the low-complexity range. Eight studies use models in the medium-complexity range (majority of them using a low input size, with the exception of CANN, BiLSTM-SEP and SEPNET), and only three employ deeper architectures with more than 10^5 trainable parameters. For context, modern large-scale ML models such as Large Language Models (LLMs; ChatGPT, Copilot, Claude, Gemini, etc.) contain billions of trainable parameters ($\geq 10^9$), while the heliophysics foundation model SuryaFM (Roy et al. 2025a) has 3.7×10^8 parameters —comparable to the EPREM-S surrogate model. This contrast highlights a clear opportunity for future research: the development and exploration of more complex, higher-capacity ML models.

Based on this complexity categorization, Figure 2a is divided into nine regions. The green regions highlight groups of models whose complexity and input size align —low complexity with small input size, medium complexity with moderate input size, and high complexity with large input size. Input size reflects the amount of information available to each model during training. Most studies (14/24) fall within these green regions (and 18/24 are within the gray dotted trend lines), supporting the intuitive assumption that larger input sizes often require deeper model architectures. One might also expect that deeper architectures would yield better performance. However, among the high-complexity models, only MEMPSEP provides predictions of geoeffective SEP events, as EPREM-S and PSPSP do not (their uniqueness is further discussed in Sections A.23 and A.24), therefore this assumption cannot be tested yet. The overwhelming majority (22/24) of models presented here predict geoeffective events, with the exception of EPREM-S and PSPSP which either do not provide a prediction (EPREM-S) or predict particles within the inner heliosphere and not at Earth (PSPSP).

In the medium-complexity and medium-input-size category, only two models appear: BiLSTM-SEP and CANN. When comparing these groups to the lower-complexity, lower-input-size models, the results in Table 5 (which will be

discussed in further detail later in the text) do not show improved performance. In fact, BiLSTM-SEP, SEPNET and MEMPSEP report lower True Skill Score (TSS) and Heidke Skill Score (HSS) scores than their simpler counterparts. Overall, no clear trend emerges linking model complexity to predictive performance. This is further illustrated in Table 5, where models are ordered by complexity, yet no increasing performance trend is observed across any of the five evaluation metrics. The white regions in the top-left and bottom-right of Figure 2a represent areas where one would not expect models to appear (and indeed none do): small input sizes do not require models with a large number of trainable parameters, and conversely, low-complexity models are generally unsuitable for processing large, information-rich datasets such as 2D inputs.

Although there is no clear evidence that increasing model complexity alone leads to improved predictive performance, this should not discourage the community from pursuing more complex architectures. As discussed in detail in Section 4.3, comparing model results at this early stage of ML-based SEP prediction is not reliable, since most reported scores correspond to different forecast windows, different prediction targets, or entirely different validation setups. What can be confirmed from our analysis is that larger input sizes (i.e., more relevant information) tend to correspond to, and often require, higher model complexity. Therefore, this study cannot conclusively determine whether more complex architectures yield better predictive performance. Nonetheless, higher-complexity models remain underexplored as seen in Table 3 and since performance improvement comes from richer input data, future progress will likely depend on combining more informative datasets with higher complexity models capable of levereging them effectively. At the current stage of the field, modelers should recognize the trade-off between training time and development difficulty on one hand, and the potential benefits of incorporating richer information on the other. As with the other qualitative characteristics in Table 3, the distribution of model architectures used across the literature is summarized in Figure 3b.

In summary, classical ML methods remain competitive for SEP prediction when data volumes are limited. Their strengths lie in robustness and interpretability, though performance improvements often require richer physical inputs rather than algorithmic complexity. Ensemble and tree-based methods provide stable performance under class imbalance and remain attractive for operational deployment due to their robustness and moderate computational requirements. Time-series architectures show promise for improving onset-time prediction by exploiting temporal evolution of solar and heliospheric measurements, though they remain constrained by limited event statistics. Deep neural architectures enable integration of complex data sources such as imagery and multivariate time-series, but their potential is currently limited by data scarcity rather than model capability. Overall, differences in predictive performance across models appear to depend more strongly on input richness and data quality than on architectural sophistication alone.

4.2. *Inputs Comparison*

The different model architectures above have been trained on inputs that vary in type, shape, size, and coverage, obtained primarily from the space missions illustrated in Figure 1. The middle columns of Table 3 (Input columns) prescribe the shape and type categories to which the inputs of each model belong. As seen in the Venn diagram of Figure 3a, most models (12/24) use some type of time series input, followed by single-point inputs (9/24) often related to a progenitor event, such as flares and CMEs (triggered prediction). Only two models, MEMPSEP and PSPSP use 2D data, such as sequences of full-disc line-of-sight magnetograms from the Michelson Doppler Imager (MDI/SOHO; Scherrer et al. 1995) and the Helioseismic and Magnetic Imager (HMI/SDO; Scherrer et al. 2012) or full-disk EUV imagery from the Atmospheric Imaging Assembly (AIA/SDO; Lemen et al. 2012), in addition to 1D quantities. Note that there are a number of studies that use more than one input, and these inputs in some cases are of different shape (TSF, MEMPSEP and PSPSP).

In terms of the type of observations used as input, ten different physical quantities have been used to train ML models. These include solar EUV imagery or magnetograms and their derivatives such as the Space-Weather HMI Active Region Patches (SHARP) and the Space-Weather MDI Active Region Patches (SMARP) data (Bobra et al. 2014, 2021). Derivatives of observations captured by the Large Angle and Spectrometric Coronagraph (LASCO; Brueckner et al. 1995) instrument, such as CME lists, have also been used to train the AA, SEP-C, MEMPSEP, ESPERTA and MS-SEP models. In terms of timeline or point data, studies have used soft X-ray, proton and electron flux, ground and space-based radio data and solar wind parameters or flare location information. The histogram of Figure 3c shows the number of models using the various input types. The overwhelming majority uses soft X-ray (15/24) and Proton Flux (13/24) data from the GOES satellites (Figure 1) because they are available in near real-time

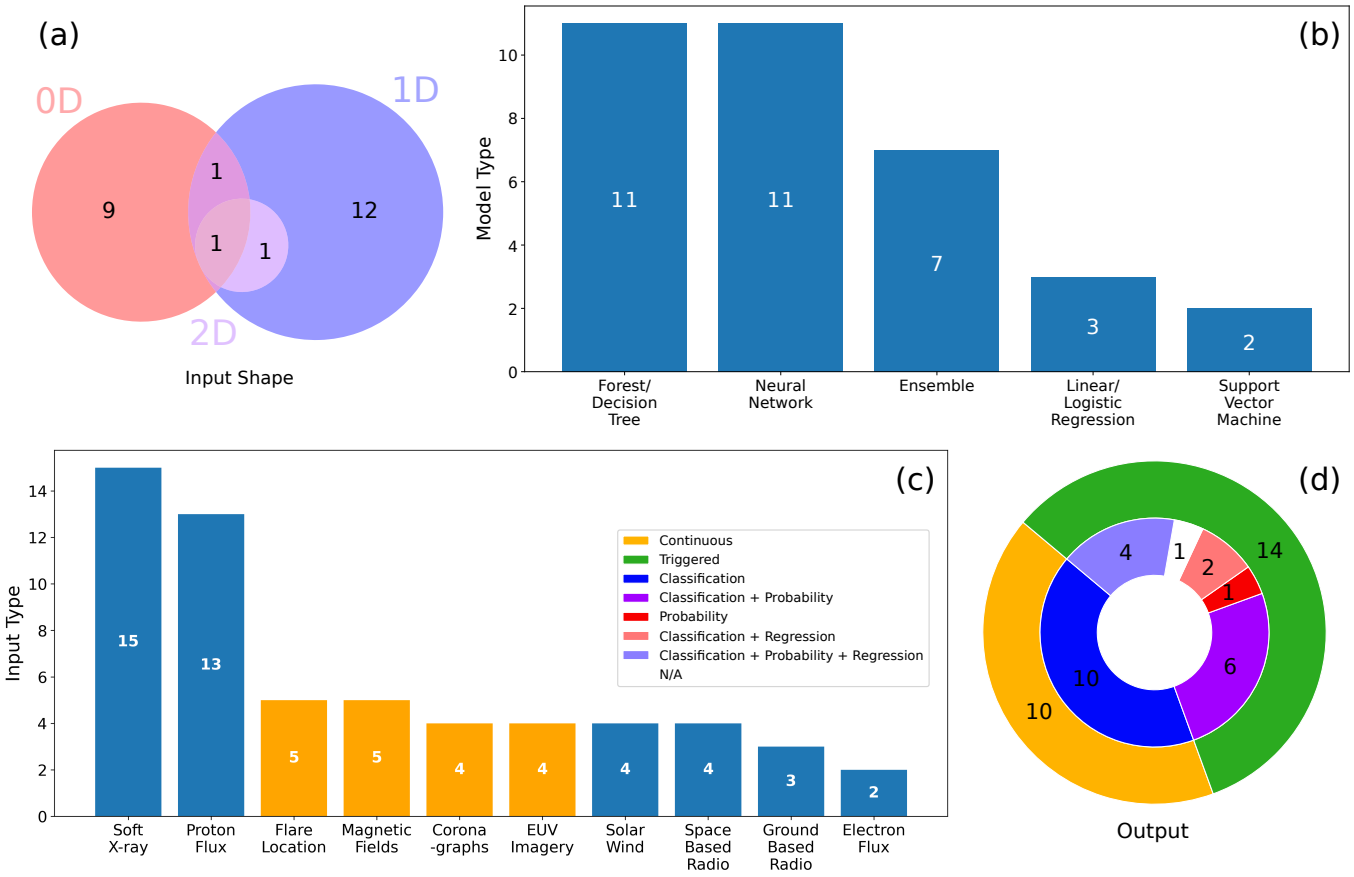


Figure 3: Summary plots of the quantitative classifications presented in Table 3. The Venn diagram at (a) and histogram at (c) summarize the *Input* columns (Shape and Type), the histogram at (b) the *Architecture* columns and the pie chart at (d) the *Output* columns (Prediction for the inner circle and Type for the outer circle). In the Input Shape and Output plots, each model is shown separately, whereas in the Input Type and Model Type histograms, methods that employ multiple types contribute to all relevant histogram bars.

(1 minute cadence[3]), extend over a long time period (since 1984) spanning multiple Solar Cycles (SCs), are obtained from well calibrated instruments —such as the X-Ray Sensor (XRS; Chamberlin et al. 2009; Woods et al. 2024)— and are used in operational decision-making by federal agencies and the commercial sector. More specifically, GOES data are used for making operational decisions at NASA and by other operational communities since they provide operationally-supported measurements which include a 24/7 support, a primary and secondary backup in case of failure, and the detectors do not saturate during high intensity periods when operations are most likely to be affected. The blue histogram bars in Figure 3c indicate in-situ measurements, while the orange bars indicate remote-sensing observations. Note that in Table 3 and Figure 3c the Magnetic Fields *Inputs* Type refers to both in-situ and remote observations although only Bi-LSTM and MEMPSEP use in-situ magnetic field observations (MEMPSEP uses both). As expected due to the nature of the SEP prediction problem, most studies use some type of in-situ measurement while a smaller number of studies complement their analysis with data captured from remote observations.

A common problem in SEP prediction, especially when using data-hungry ML models, is the absence of positive events (SEP occurrences)– what is commonly known as data imbalance. More specifically, flare-based studies encounter the flare imbalance problem where the overwhelming majority of flares do not produce SEPs (negative), while for models that use time-series the majority of their data is labeled as "quiet times" due to the rarity of significant space weather events. Here, we express the imbalance in each model as a qualitative parameter in Tables 6-29. The imbalance parameter is defined as the ratio of positive events over the total number of events, therefore, it is always an integer

[3] https://www.swpc.noaa.gov/products/goes-x-ray-flux

between 0 and 1. Figure 2b shows the diversity of each model (number of total samples used; the number of SEP and non-SEP occurrences together) against the imbalance parameter. As expected, models trained on larger number of samples (more diverse) are highly imbalanced (their imbalance ratio is lower), while models trained on a smaller number of samples are less imbalanced (imbalance ratios closer to 1). EPREM-S is again a special case where the model sees only SEP occurrences.

Interestingly, two models that deviate strongly from the trend in Figure 2b (the red dotted best-fit line) —SEP-C and RH— are also ones that report the highest Probability of Detection (POD ≥ 0.9) and TSS (≥ 0.9), but at the same time the lowest HSS (≤ 0.25), as shown in Table 5 and will be further discussed in the text that follows. Both studies train their models on a very small number of positive events compared to other works with a similar total number of samples (e.g., BiLSTM-SEP and UNSPELL). This type of performance is common in highly imbalanced scenarios in which many correct negatives result in high POD but do not necessarily represent skill in predicting rare events (Peirce 1884; Whitman et al. 2026). In rare-event settings, TSS can remain high when most positive events are captured (high POD), because the False Alarm Rate (FAR) is also very high (for SEP-C, a FAR of 0.88 is reported), making the models non-useful for operational predictions, where lots of false alarms become an inconvenience. This might also occur because correct rejections dominate the confusion matrix, causing TSS to track POD closely. Therefore, having the highest TSS and POD does not imply that SEP-C and RH are the best-performing models, nor does it suggest that studies should avoid training on larger sets of positive events. Rather, it demonstrates that high scores can be misleading when viewed in isolation —especially in rare-event regimes— and must be interpreted alongside other metrics such as HSS, which in this case clearly indicates limited overall skill. Recall that multiplying the model imbalance (Figure 2b; y-axis) with the diversity (x-axis) yields the number of positive events available for training each model.

In future studies, data augmentation and models that can produce realistic SEP distributions should be explored as a remedy to imbalance. Imbalance mitigation through data augmentation (Bahri et al. 2023), has been applied to SEP prediction only in the study by Hosseinzadeh et al. (2024a) (see Table 2). The results of the relevant model (TSF) appear promising: the model reports high TSS and HSS values of 0.80 and 0.90, respectively. However, this is only a single data point, so conclusions should be made with caution until more studies reproduce similar results. It should also be noted that even with augmentation, TSF still exhibits relatively low imbalance compared to other studies with a similar number of samples. Nevertheless, the community should explore this direction further, as data augmentation has already shown performance improvements in flare prediction tasks (Li et al. 2025; Wen & Angryk 2024; Grim & Gradvohl 2024). Beyond this, no major trends or correlations between dataset diversity, imbalance, and the reported model performance can be identified from the existing literature, mainly due to the difficulty in comparing fairly the models' results.

Lastly, an important input parameter is the input size, which in this discussion is particularly relevant because larger inputs generally require more complex models to process them effectively. For this reason, input size is plotted against model complexity in Figure 2a. The colors of the points indicate the data dimensionality used by each model (0D, 1D, 0D+1D, or 2D), following the classification in Table 3. As expected, models using 2D data exhibit large input sizes and correspondingly rely on more complex architectures. Since input size reflects the amount of information processed per sample, it is reasonable to hypothesize that models with larger input sizes should achieve better performance. However, the current results presented in Table 5 do not support this expectation. Models in the mid-range of input sizes (BiLSTM-SEP, CANN, CART, and STSF) report some of the lowest POD values (0.73 for CART and 0.61 for BiLSTM-SEP, where available), yet at the same time achieve high F1 scores (e.g., 0.82 for CART) and notably low FAR (0.09 for BiLSTM-SEP). Attention should also be given to models with small input sizes but medium complexity—such as TS-HOG-TB, TSF, UDM, SPRINTS, and UNSPELL. Despite using relatively limited inputs, these models train more parameters and achieve performance comparable to models with substantially larger input sizes. These architectures (particularly NNs and ensemble forests) appear capable of extracting meaningful information even from smaller datasets and warrant further exploration. Nevertheless, due to the inconsistent use of validation metrics across studies, it remains difficult to draw strong conclusions about the relationship between input size, complexity, and performance at this stage. This difficulty is discussed extensively in Section 4.3.

This review covers studies that were published beginning in 2011 (Núñez 2011), during the ascending phase of SC 24. Since then, twenty three more models have been published using the more recent data, up to the maximum of SC 25. Therefore, different models have used data from different periods of solar activity (Figure 4). Most of the models discussed here use data from SCs 23 and 24, while nine studies have used historical data from SC 22. Only

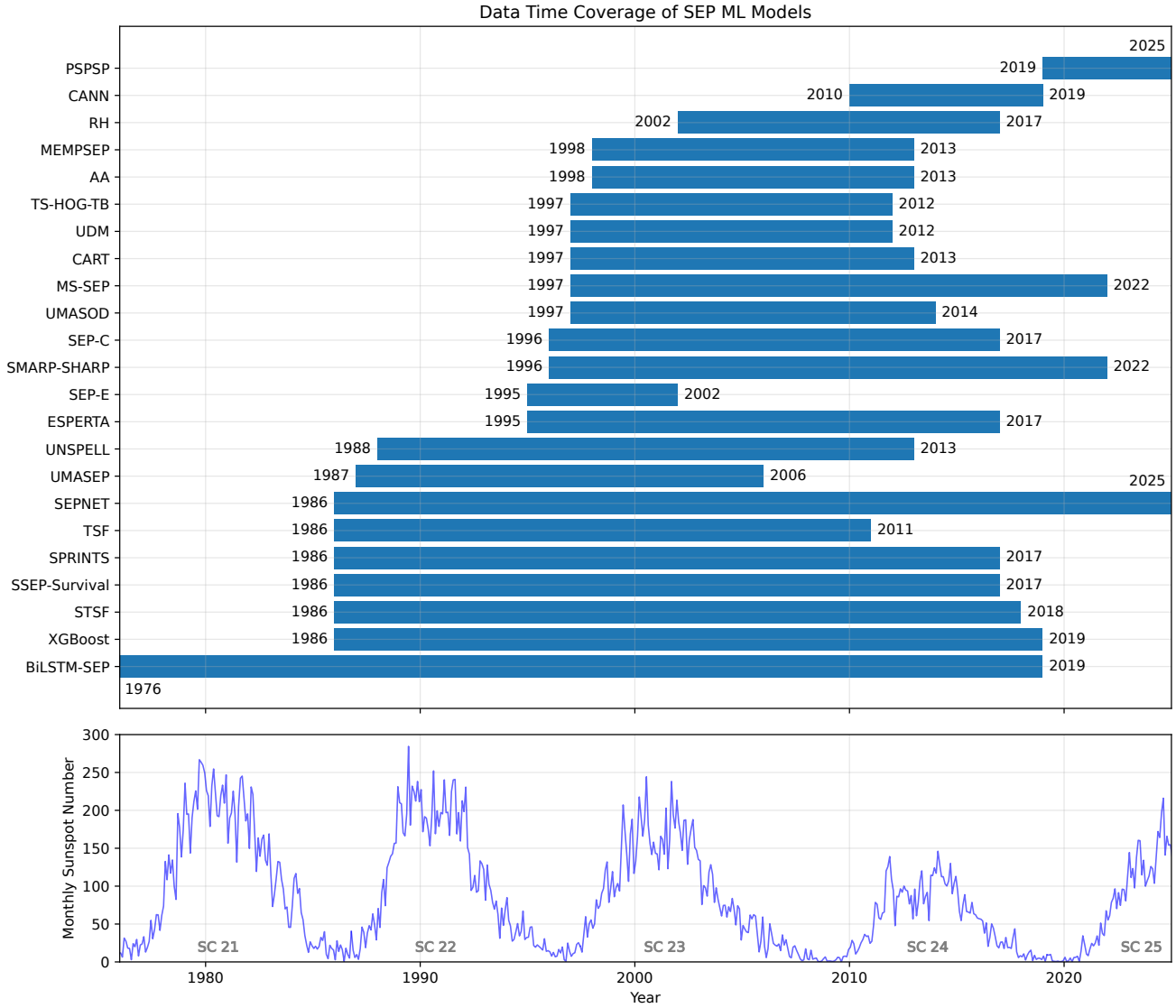


Figure 4: Time coverage of the reviewed SEP prediction models (start and end years, per study) shown alongside the monthly sunspot number, illustrating how each model's training and evaluation data align with the SC variability.

the BiLSTM-SEP model includes data from SC 21. Note that although most publications cited in this work describe models that appear static in time, there are models such as UMASEP and SEPNET that are being continuously updated and their data now spans up to the most recent data of SC 25, similar to the PSPSP model. Since most observatories (Figure 1) providing data for these models remain operational —although some, such as ACE, now produce data of questionable quality (Regnault et al. 2020)— the majority of models could in principle be updated with the most recent observations, but doing so remains at the discretion of the developers.

4.3. *Outputs and Testing*

Although all models described in this manuscript have been developed for achieving the same goal —predicting SEP events, their predictions are different, given that they have different inputs and are developed to forecast different output quantities. The differences between the outputs are captured on the rightmost columns of Table 3 (Output columns) where the predictions are classified in two groups: the *Prediction* group (Classification, Probability, and Regression) and the *Type* group (Triggered or Continuous). More information on the definitions of these categorizations is included in Section 2 and the questionnaire in Appendix C. The prediction *Types* are mutually exclusive: a triggered model (in which a prediction is issued based on an event such as a flare or CME; post-eruptive) cannot be continuous (in which

predictions are issued continuously without dependence on an event; pre-eruptive), and vice versa. On the other hand, the *Prediction* categorization can be often mutually inclusive, meaning that a model that outputs probabilities can also provide a classification given a defined threshold.

The nested pie chart of Figure 3d summarizes the nature of the predictions provided by the models in this review. There is an almost even split between continuous (10/24) and triggered (14/24) predictions (outer pie chart). The inner pie chart shows that the overwhelming majority (22/24) of models output an event classification, with ten of them also providing a probability. All the regression models use a threshold for an SEP event in order to provide a classification. Only a minority of models do not provide an event classification (SSEP and EPREM-S). For instance, EPREM-S is a NN trained on synthetic data, not built to predict the occurrence of SEP events (though it has the capability), the SSEP model outputs a function that indicates the probability that an SEP event has not occurred at a given, post-flare, time. Another difference among the model outputs is the forecast window, which is plotted against the sample coverage Figure 2c. Although we expect that smaller forecast windows should produce better results (Table 5), even this expectation is not obvious here, underlying once again the difficulty of comparing between the different model results in this stage of the community's efforts.

Since most models provide some type of binary classification, whether this is an "All Clear" or a binary yes/no for SEP events ($>$ 10 MeV), the community has used a wide variety of evaluation metrics that capture different aspects of the predictive capabilities of each model. Table 4 presents a summary of all the metrics that have been used by the community to validate their ML models. More specifically, the different columns represent the eight most used metrics: the True Skill Score (TSS; Doswell et al. 1990), the Heidke Skill Score (HSS), the Accuracy (ACC), the Probability of Detection (POD or Recall; Wehling et al. 2011), the False Alarm Rate (FAR; Macmillan & Kaplan 1985), the F1 score (F1; Lipton et al. 2014), the Precision (Prec) and the Area Under the Curve (AUC; Muschelli III 2020). Here we need to note that although most of the metrics are derivatives of a contingency table, the AUC is different in that it captures the best performing threshold introduced in order to turn an output probability into a binary classification. The last column of Table 4 presents other, less popular metrics used for model evaluation. The FAR measures the fraction of all non-events that were incorrectly predicted as events (False Alarms/Total Non-Events). On the other hand, the False Alarm Ratio (FAR*) measures the fraction of all "yes" forecasts that were wrong (False Alarms/Total Forecasted Events). Note that herein, FAR indicates the False Alarm Rate whereas the FAR* is the False Alarm Ratio. In practice, the FAR* is often the more challenging metric to optimize, as it penalizes every incorrect 'yes' forecast directly and therefore reflects the forecaster's precision more strictly than the FAR. The last column of Table 4 presents a list of other, less utilized by the community metrics, such as the Pearson Correlation Coefficient (PCC; Benesty et al. 2009; Kasapis et al. 2023b), the Root Mean Squared Error (RMSE; Hodson 2022), the R^2 Score (Ash & Shwartz 1999), the Kaplan-Meier estimate (KM; Goel et al. 2010), Balanced Accuracy (BA) and others. A list of acronyms, which includes the aforementioned metrics, is available in Appendix D.

Five of the aforementioned metrics —namely the POD, TSS, HSS, F1 and FAR— have been used by the majority of the studies. To facilitate comparison across models, Table 5 summarizes the values of these metrics as reported in the respective publications. However, several difficulties arise when attempting to compare results between studies. Most works generally follow the NOAA SEP event definitions[4], in which proton event alerts are issued for several thresholds and two particle-energy levels, with the $\geq$ 10 MeV channel aligned with the NOAA Solar Radiation Storm S-scale thresholds —10, 100, 1000, 10000, and 100000 pfu (S1-S5 thresholds)— and the $\geq$ 100 MeV channel being based on a single threshold of 1 pfu. Nevertheless, the specific thresholds adopted by each study vary. For example, models such as CART, TSF, UDM, and TS-HOG-TB use a $\geq$ 100 MeV threshold, whereas most studies classify events using the $\geq$ 10 MeV criterion. Additionally, some studies do not employ the NOAA definition at all. UNSPELL, for instance, uses the European Space Agency (ESA) Solar Energetic Particle Environment Modeling (SEPEM) reference event list[5], a standardized catalog of solar proton events (SPEs). Other approaches diverge for methodological reasons: EPREM-S does not perform event prediction, and PSPSP defines a mission-specific threshold tailored to Parker Solar Probe measurements, since it does not aim to predict geoeffective events. When the SEP event definition differs between studies, and when training and testing conditions are not identical, a fair comparison of the reported results becomes impossible.

[4] https://www.swpc.noaa.gov/products/goes-proton-flux

[5] http://sepem.eu/help/event_ref.html

Table 4: Summary of metrics used to validate the SEP ML models' outputs.

Model	TSS	HSS	ACC	POD	FAR	F1	Prec	AUC	Other Metrics
A.1 XGBoost	✗	✗		✗					
A.2 STSF	✗	✗				✗		✗	MCC, GSS
A.3 SMARP-SHARP	✗	✗	✗	✗	✗	✗			
A.4 AA	✗	✗		✗	✗	✗			
A.5 ESPERTA				✗	✗	✗			CSI
A.6 UMASEP				✗	✗				
A.7 UMASOD				✗	✗			✗	
A.8 MS-SEP	✗	✗		✗	✗	✗			
A.9 CART			✗	✗		✗	✗	✗	
A.10 RH	✗	✗	✗	✗			✗	✗	BA
A.11 SSEP									KM Est, Cox PH & Param. Models
A.12 SEP-C	✗	✗		✗	✗	✗	✗		
A.13 CANN	✗	✗						✗	
A.14 SEP-E				✗		✗	✗		MAE
A.15 SPRINTS		✗		✗	✗				
A.16 TSF			✗			✗			
A.17 UDM	✗	✗	✗	✗		✗	✗		
A.18 UNSPELL	✗			✗	✗			✗	
A.19 TS-HOG-TB	✗			✗		✗	✗		
A.20 SEPNET	✗	✗	✗	✗	✗	✗		✗	
A.21 BiLSTM-SEP	✗	✗	✗	✗	✗		✗		CSI
A.22 MEMPSEP	✗	✗	✗	✗				✗	R^2, RMSE, PCC, BS, ECE, FPR
A.22 PSPSP	✗	✗	✗	✗	✗	✗	✗		
A.24 EPREM-S									Deep Ensemble (see Appendix A.24)
Total	15	14	9	19	12	13	8	8	

As is evident in Table 5, it is difficult to compare even studies that adopt a common SEP event definition because they often do not evaluate their performance using the same set of metrics. Although having multiple common metrics is important and necessary, as comparison based on a single metric is biased, it is often insufficient for a fair comparison of model predictions. For instance, although the MS-SEP, SMARP, and AA models share three metrics in common (TSS, HSS, and F1), they employ different forecast windows of 5, 14, and 24 hours, respectively. Figure 2 also shows that both the forecast windows and the corresponding input-sample coverage vary widely across studies, ranging from just a few hours to as long as four days for models such as SPRINTS. A substantial number of studies do not report a forecast window at all, either because they do not issue a prediction in the sense of an advance alert, or because they do not provide a geoeffective event prediction (as is the case for EPREM-S or SSEP). There are also differences in the physical targets and prediction objectives of the models. Forecast horizons and predicted quantities vary considerably. For example, some models attempt to determine whether a space weather event (flare, CME, etc.) will produce an SEP (models marked as Triggered in Table 3), whereas others provide all-clear predictions (CANN) or estimate particle fluxes like BiLSTM-SEP (which may then be reduced to a binary outcome). Regardless of these methodological differences, most studies ultimately reduce their output to a binary yes/no prediction and report a set of common metrics, which are summarized in Table 5.

Despite these inconsistencies, a general comparison across studies is still possible using the metrics in Table 5. For example, the SMARP–SHARP model shows relatively low performance (TSS: 0.39 ± 0.19, HSS: 0.37 ± 0.19). This suggests that relying on a low-complexity model (SVM) together with only point-data (0D), such as features derived from magnetogram data, likely provides insufficient information for reliable SEP prediction. Although its TSS and HSS values are modest, the model achieves a reasonably low FAR* (0.30 ± 0.14), indicating that while it may struggle to correctly predict positive events, it does not tend to over-predict them —making it a conservative model.

Table 5: Summary of the validation results for the 24 SEP ML models, as obtained from the relevant publications linked in Table 1. The FAR results with asterisk* indicate the False Alarm Ratio, not Rate. As mentioned in the text, direct comparison between models, based on the validation measures listed, should be done with caution as the models have been applied with different training and testing setups. EPREM-S adopted a deep learning approach for determining uncertainties of the EPREM-S outputs rather than applying traditional skill scores (MSE of 0.07 was obtained). SSEP did not use for validation any of the five listed metrics, but evaluated performance using the Kaplan–Meier (KM) Estimate and the Cox Proportional Hazards (Cox PH). For studies that have quantified their results' uncertainties, metrics are shown as the mean ± one standard deviation across k cross-validation folds.

Model	POD	TSS	HSS	F1	FAR	SEPVAL
A.1 XGBoost	0.81	0.72 ± 0.02	0.42			
A.2 STSF		0.85	0.88			
A.3 SMARP-SHARP		0.39 ± 0.19	0.37 ± 0.19	0.67 ± 0.11	$0.30 \pm 0.14^*$	
A.4 AA	0.76 ± 0.06	0.75 ± 0.05	0.69 ± 0.04	0.70 ± 0.04	0.34 ± 0.10	Yes
A.5 ESPERTA	0.88			0.77	0.32	
A.6 UMASEP	0.81				0.34	Yes
A.7 UMASOD	0.85				0.85^*	
A.8 MS-SEP	0.85 ± 0.08	0.78 ± 0.07	0.71 ± 0.03	0.75 ± 0.03	0.31 ± 0.08	
A.9 CART	0.73			0.82		
A.10 RH	0.96 ± 0.00	0.94 ± 0.01	0.17 ± 0.01			
A.11 SSEP						
A.12 SEP-C	0.92	0.91 ± 0.04	0.25 ± 0.06	0.25 ± 0.06	0.88^*	
A.13 CANN		0.82 ± 0.01	0.38 ± 0.03			
A.14 SEP-E	0.70			0.76		
A.15 SPRINTS	0.86				0.37	Yes
A.16 TSF		0.80	0.90			
A.17 UDM				0.79		
A.18 UNSPELL	0.86	0.78			0.08	Yes
A.19 TS-HOG-TB	0.81			0.80		
A.20 SEPNET	0.64	0.43	0.42	0.71	0.23^*	Yes
A.21 BiLSTM-SEP	0.62	0.53	0.73		0.09	
A.22 MEMPSEP	0.83	0.63	0.60			Yes
A.22 PSPSP	0.70 ± 0.17	0.43 ± 0.14	0.31 ± 0.12	0.46 ± 0.10	0.64 ± 0.10	
A.24 EPREM-S	Uncertainty estimation as explained in the caption and Appendix A.24.					

Nevertheless, such low-complexity models (e.g., XGBoost, STSF, AA, etc.) remain valuable: their interpretability helps identify which parameters in datasets are informative for the SEP prediction task. The AA model, which is similar to the SHARP–SMARP in terms of overall methodological simplicity but employs soft X-ray measurements and coronagraph data, performs substantially better (TSS: 0.75 ± 0.05, HSS: 0.69 ± 0.04). This improvement can be attributed to the richer physical information content available in these additional data sources. At the other end of the spectrum, MEMPSEP uses the most diverse physical dataset (Magnetic Fields, soft X-ray, Electron Flux, EUV Imagery, Coronagraphs, Space-Based Radio and Solar Wind) and employs a high-complexity ML model (6,092,617 trainable parameters); however, despite its relatively short forecast window (6 hours), it performs well in POD (0.83) but not as well in TSS or HSS (0.63 and 0.60). RH and SEP-C have unusually high POD values (0.96 ± 0.0 and

0.92). However, this is paired with very high FAR* values (0.882; RH does not explicitly report FAR, but its low HSS suggests a similarly high FAR*). In contrast, the BiLSTM-SEP model achieves a lower POD (0.62), but also a very low FAR (0.09), a combination widely regarded as a strong indicator of a reliable SEP prediction model. Ideally, POD should be high while FAR remains low, but in practice these metrics compete, and modelers navigate the trade-off between them.

Despite limitations noted above, certain patterns are already observable. For example, models achieving high detection rates often suffer from elevated FAR. This is directly related to the imbalanced problem at hand, as analytically presented in Stumpo et al. (2021) and in Lavasa et al. (2021), with a typical FAR ranging between 0.25-0.45 (see the discussion in Papaioannou et al. 2025). Operational forecasting requires balancing competing objectives depending on user needs. For example, aviation and astronaut safety operations often prioritize minimizing false alarms, whereas scientific monitoring efforts may tolerate higher false alarms in order to avoid missed events. Despite their elevated FAR or reduced POD, ML models can still be operationally valuable because they often provide substantially greater lead times.

Because inconsistent SEP event definitions and non-uniform training and testing conditions make fair comparison across studies impossible, a community-standardized validation framework such as the SEPVAL challenge is essential. The Solar Energetic Particle Model Validation Challenge (SEPVAL), organized by the Space Radiation Analysis Group (SRAG) at the NASA Johnson Space Center (JSC) through NASA's Integrated Solar Energetic Proton Event Alert/Warning System (ISEP) collaboration, is an ongoing community-wide effort designed to develop a generalized framework and prescribed methodology to assess the performance of operational and research SEP forecasting models, using standardized event lists, metrics, and evaluation procedures. SEPVAL brings together model developers and space weather end-users to compare SEP predictions against a consistent observational dataset. By providing common input parameters, clearly defined validation periods, and transparent scoring methodologies, SEPVAL aims to identify model strengths and weaknesses, quantify forecast skill, and promote best practices in SEP prediction. SEPVAL provides a platform for more consistent, streamlined, and fair ML model comparisons than what can be achieved through review efforts like this manuscript, effectively addressing the comparison challenges outlined in this work.

To date, 24 SEP models of all types have participated in SEPVAL, including 6 ML models (marked in Table 5). A description of the validation methodology developed through SEPVAL and a summary of model performance for participating models is published by Whitman et al. (2026). The SEPVAL challenge has so far focused on two operational thresholds important to SRAG, the $\geq$ 10 MeV that exceeds 10 pfu and the $\geq$ 100 MeV that exceeds 1 pfu, for an approximately balanced set of challenge periods comprised of 33 SEP events and 30 non-event periods. All periods are associated with fast CMEs and strong flares, relevant to SRAG operations. Modelers have submitted predictions of all types, including probability of occurrence, binary all clear, peak intensity, fluence, and full-time profiles. Each participating model has a unique set of inputs and outputs which, as has been discussed above, makes it difficult to compare models directly. However, SEPVAL organizers made the decision that the models could be viewed as a group or ensemble and that a mean and median performance derived from this group could be used as a meaningful definition of the state-of-the-art of SEP model performance. In Whitman et al. (2026), the participating SEPVAL models were divided into two groups, pre-eruptive (continuous) and post-eruptive (triggered), and median and top quartile scores were reported for selected metrics for probability, all clear, and peak intensity. Using these metrics as a reference, SEP models may evaluate their predictions for the set of 63 SEPVAL challenge periods and determine whether their performance exceeds the median with the goal of achieving top quartile performance.

Whitman et al. (2026) calculated the state-of-the-art median and the top quartile performance for 10 models predicting in real-time on the SEP Scoreboards[6], representing a realistic operational forecasting scenario that includes real-time data latency, data gaps, human-in-the-loop analyses, and true highly-imbalanced climatology. While it may be difficult for models to make a direct comparison to these metrics, they indicate more realistic model performance in an operational setting. If model developers were to run their models in real time for an extended period or perform a simulated real time evaluation, the scores derived from the SEP Scoreboards are most appropriate for comparison.

It should be understood that reliable prediction of SEP events using ML is a complex task. Not only are the generation, transport, and propagation of energetic particles from the photosphere to Earth still active areas of research, but the amount of relevant heliophysics data, although it has increased over the past decades, is still very sparse compared to other problems for which ML has been applied and now outperforms classical methods. It should

[6] https://ccmc.gsfc.nasa.gov/scoreboards/sep/

also be noted that the studies presented in this review represent the very first efforts of the heliophysics community to predict SEP events using the new tool-set provided by ML. Although the community is not yet able to make reliable comparisons between results, nor it is at a stage where operationally robust SEP predictions can be produced, the insights summarized in this paper can help outline several promising paths for future research, which are discussed in the next section.

5. PATHS FOR FUTURE RESEARCH

Here, some additional observations about the model results can be made, along with recommendations for future research. First, as seen in Table 5, many studies do not quantify the uncertainty of their models' performance (e.g., by reporting forecast standard deviations). This implies that either k-fold cross-validation was not performed, or the corresponding variability was not reported. Repeated cross-validation with different random splits reduces the randomness associated with any single train–test split and yields a more statistically robust and trustworthy estimate of model performance. Therefore, it is strongly recommended that future studies employ repeated k-fold cross-validation and report at least the standard deviation of the model results over the ensemble of k-fold splits. Quantification of uncertainty should not be overlooked in ML studies in heliophysics (Keegan et al. 2025).

Furthermore, different studies use different sets of evaluation metrics —or sometimes only one or two metrics— making consistent comparison extremely difficult. We recommend that all future studies that ultimately reduce their predictions to a binary yes/no SEP classification, report all five commonly used metrics listed in Table 5, together with their standard deviations. POD and FAR provide a complementary picture of event detection performance: POD captures the ability of a model to correctly identify events (sensitivity), while FAR quantifies the frequency of false positives. Together, these two metrics offer a holistic understanding of operational reliability. FAR is a critical metric for operational forecasting; Núñez (2011) emphasizes its importance extensively, noting that an effective prediction system should aim to maximize POD while simultaneously minimizing FAR. Meanwhile, TSS is widely used because it adjusts for class imbalance, incorporates both false positives and false negatives, and provides a normalized measure of predictive skill, offering additional insight beyond raw detection rates.

In addition, it would be beneficial for the community to streamline, to the extent possible, the choice of forecast window. As shown in Figure 2c, the majority of studies aim to predict SEP events within a 24-hour window. A 24-hour prediction is operationally meaningful for mitigating radiation risks from geoeffective events. Adoption of standarized forecast windows greatly enhances comparability between studies. It is also informative to examine model performance across multiple forecast horizons —a strategy adopted by studies such as BiLSTM-SEP, SEP-E and MEMPSEP. It is therefore recommended that future work explore at least three representative forecast windows, such as 6 hours, 12 hours, and 24 hours. Studying how performance varies with forecast horizon provides insight into the temporal limitations of the model, the persistence of precursors, and the time window in which predictions are most reliable for practical operational use.

Lastly, as mentioned previously, the majority of studies define an SEP event using the ≥ 10 MeV threshold introduced by NOAA. While this definition is widely used, for operational and geoeffective SEP prediction it is important to recognize that higher-energy particles are often the most critical. In particular, ≥ 100 MeV proton enhancements are especially critical for operational space-weather decision making, as they directly affect astronaut safety, can increase radiation exposure for high-altitude aviation, and can disrupt spacecraft operations through communication and navigation disturbances (radio storms) as well as increased atmospheric drag due to upper-atmosphere inflation. Future work should therefore prioritize forecasting —rather than only predicting— particularly with respect to SEP onset time, peak intensity, and total event fluence at locations of interest, other than Earth (e.g., cislunar space or Mars), in order to support NASA exploration activities and other deep-space operational needs. Among the methods reviewed in this study, only the PSPSP model offers some insights about SEP prediction in areas other than near-Earth. Even proof-of-concept ML studies should aim to design their models and outputs as close to operational requirements as possible. Model developers should also explicitly state which operational organizations or user groups their models are intended to support.

In addition to the above considerations, it is crucial for future SEP prediction studies to address how their models would operate in real-time settings. A key first step is to identify the intended end user and understand what they require from an SEP prediction system. Different organizations —such as the SRAG, the Space Weather Prediction Center (SWPC), aviation radiation authorities, satellite operators, or mission planners— have distinct operational needs, and the design of a prediction model should reflect these needs. For example, for the NASA JSC SRAG and

agencies concerned with aviation radiation control, reducing false alarms is of the highest priority, as unwarranted alerts carry significant operational cost. For these users, prediction of ≥ 10 MeV events is useful, but accurate prediction of ≥ 100 MeV events is especially valuable, given their stronger radiation impact and operational relevance.

Model developers must also consider how the end user intends to use the predictions, in what environment, and under what constraints, following the Research-to-Operations (R2O[7]) framework. It is important to distinguish between a model that performs well in an operational setting and one that merely runs in real time. Operational readiness requires that the system handle errors, data gaps, and degraded inputs gracefully. If a model fails —or is unable to provide predictions— when data are missing, delayed, or corrupted, then it is not operational in practice, regardless of its offline performance. This relates to the broader concept of system robustness and the SWPC readiness levels, where performance testing in real-time conditions is only one (albeit important) component of operational validation.

At present, most ML-based SEP prediction methods are not operational for two main reasons. First, some rely primarily on data that are not available in real time for operational use. For example, data products such as the SMARP-SHARP dataset by Kosovich et al. (2024) (used by the SHARP-SMARP and CANN models), synthetic data (EPREM-S), and even certain high-energy proton measurements (PSPSP) are often unavailable because they are processed manually. Second, although their input data can be accessed in near real-time (all models that use GOES data as input), the readiness of the models themselves is limited. Few studies (e.g., UMASEP) provide the software infrastructure required to deploy their models online, maintain continuous ingestion of real-time measurements, and deliver predictions continuously and without interruptions (real-time space-based observations quite often have data gaps). For a model to be considered operational, it must be supported by automated pipelines, error-handling systems, and ability to run continuously without manual intervention. Complementing the recommendations above, future studies should therefore explicitly address both data availability and model readiness when developing ML-based SEP forecasting systems.

In regard to the inputs, majority of the models (15/24) use soft X-ray measurements from the GOES satellites, due to their low down-link time, ease of use (timelines) and direct relation to space weather activity. On the other hand magnetic field inputs, coronagraphs and electron flux, although they provide meaningful signatures for space weather events, they have been underutilized as they often require more processing and larger models. The community should move towards the use of larger, more complex models (the majority of models in this study have less than 1,000 trainable parameters) that are trained on multiple physical parameters and data streams. As modern ML research has shown, larger models trained on broader and richer datasets consistently outperform smaller ones, especially for complex predictive tasks (see text prediction; e.g., ChatGPT, Claude, Copilot and others). Therefore, while small datasets and lightweight models may be suitable for proof-of-concept studies, operational SEP forecasting systems must leverage as many physical parameters and data streams as possible to achieve reliable predictions.

Future research should also experiment with new inputs and improved datasets or observations. Since February 1st, 2026, the NASA Interstellar Mapping and Acceleration Probe (IMAP; McComas et al. 2018, 2025), located at the L1 point, provides new coordinated and comprehensive observations of the inner heliosphere. In addition to science observations, five in situ instruments on IMAP also make low-latency measurements (e.g. magnetic field, solar wind electrons and protons, energetic protons and electrons) of relevance to space weather operational forecasting. Through the IMAP Active Link for Real-Time (I-ALiRT; Lee et al. 2025) space weather data system, these measurements are continuously telemetered in near real-time to Earth. I-ALiRT is based on the data system used for ACE, and therefore it provides similar space weather data products at significantly enhanced cadences, in addition to the new parameters offered. Similarly, new missions such as NOAA's Space weather Observations at L1 to Advance Readiness - 1 (SOLAR-1; formerly called the Space Weather Follow On – Lagrange 1, or SWFO-L1), and the Polarimeter to Unify the Corona and Heliosphere (PUNCH; DeForest et al. 2022) already provide new data relevant to space weather events prediction.

The majority of geoeffective SEPs originate from solar surface locations that are visible from the Earth-Sun line, where GOES and SDO are stationed. However, Richardson et al. (2014) found that on average one quarter of significant SEPs in geospace originate from source locations behind the western limb of the Sun. Most ML models use GOES X-rays or SDO magnetograph observations. While X-ray flares can still be detected, at least partially, if they occur near the limb on the far side, their brightness is reduced as compared to the same event occurring on the near side. Flares occurring farther behind the limb would be entirely missed by GOES. Also critical, currently solar

[7] https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/03/03-2022-Space-Weather-R2O2R-Framework.pdf

magnetograph measurements cover under 40% of the visible solar disk. They become inaccurate near the solar limb, with signal degradation starting already at 60° from the observer's subsolar point. This degradation affects forecasts of a substantial fraction of SEP events in geospace, given that Earth is, on average, magnetically connected to solar longitudes near $W60°$. The combined gaps in accurate observational coverage limit the POD of almost all current forecasting models listed in Whitman et al. (2023) and in this work. In effect, they are an obstacle for many models to become fully operational.

The most effective mitigation of the critical observational gaps near and behind the Sun's western limb is a mission to Earth-Sun Lagrangian point 4 (Bemporad 2021; Posner et al. 2021; Cho et al. 2023). The L4 point is gravitationally stable and is located directly over the magnetic footpoint of Earth, therefore allowing for excellent coverage in X-rays and of magnetographic observations in support of SEP forecasting models. X-ray flares can be fully observed from the Eastern limb of the Sun as viewed from Earth to $W150°$, covering essentially all "missed" events identified in Richardson et al. (2014). Magnetograph coverage will extend from $E60°$ to up to 30° behind the western limb of the Sun, with the opportunity of stereoscopic views of ARs traversing from the central meridian to $W60°$ if both geospace and L4 locations are equipped with such instrumentation. Thus, a space weather mission to L4 with real-time downlink can elevate SEP forecasting models into operational models for geospace.

In summary, considering the analysis presented in this work and the effort made to compare the community's SEP prediction ML models, it is highly recommended, if the problem setting allows, that future studies follow the list of "good practices" presented bellow:

- **Use of common validation metrics:** A difficulty encountered by the authors of this work while trying to compare the different models, is that modelers use different evaluation metrics for testing. It is recommended that future works will use at least the five metrics outlined in Table 5, as they are deemed most useful and are used regularly by the community. This suit of metrics capture well, from different angles, the performance of the model, and adapting to this common validation setup makes one's work much easier to compare to those of the community.
- **Uncertainty quantification:** ML models often tend to overfit on the data they have seen during training and can appear overly confident for some of their predictions. The best way to assess the stability of an ML model, mitigate overconfidence and prove its robust performance, is to cross-evaluate and estimate uncertainty. It is recommended that future research report appropriate measures of uncertainty quantification for the method and the type of task (e.g., regression or classification), such as prediction variance, entropic measures, prediction intervals, etc. Ideally, the reported measures should integrate both aleatoric (due to inherent randomness in the data) and epistemic (due to scarce data or knowledge gaps) uncertainties.
- **Use of common prediction window:** Another difficulty that one will encounter when trying to compare between SEP prediction works, is that oftentimes the scores reported are for different forecasting setups, prediction windows and average waiting times. Figure 2c shows that a considerable number of studies have used a 24-hour forecasting window. Many studies that report results for a 24-hour forecasting window, also evaluate model performance for windows of different length as well, as all forecast windows are operationally useful. Smaller windows often produce more confident predictions whereas larger windows allow for more reaction time. Therefore, to make one's work easier to compare with the community's, it is recommended that studies assess and report model performance for time windows of different lengths along with the 1-day-ahead (24 hours) prediction.
- **Deeper models on larger datasets:** It has been a decade since the inception of ResNets (He et al. 2016), and eight years since Transformers were introduced (Vaswani et al. 2017). In industry, models with billions of trainable parameters are being trained on terabytes worth of data. With the commissioning of new heliophysics missions, large amounts of new space weather data will be available to the community. In light of these advances, it is recommended that in the future, researchers train larger models with more data than the current state of the art (Figure 2a). This will allow models to leverage potentially useful physical information in the augmented datasets, hopefully leading to more accurate forecasts.
- **Operational-as-possible validation:** Modelers often choose validation setups that aim to favor their method and make it appear more competitive in the field. However, a fair comparison of results across different studies should take in consideration the operational level of the model. It is therefore recommended that modelers should aim for operational validation settings and model development. A model that performs very well in a simulated environment but fails to perform in real life is less useful than one which can provide real-time predictions but registers worse performance metrics.

- **SEPVAL validation on common events:** SEPVAL is a common validation scheme for operational SEP forecasts, led by the NASA JSC SRAG and the CCMC. It is highly recommended that future works build their models around SEPVAL, in order to validate on the same SEP events as other SEPVAL contributors. SEPVAL is currently the only common validation scheme that exists in the community and ensures comparison between SEP prediction models.
- **NASA Open Science:** Open sharing of data, information, and knowledge within the scientific community and the wider public accelerates scientific research and understanding. It is recommended that future authors of SEP prediction based on ML approaches, make their data and algorithms publicly available and their experiments easy to reproduce. The best way to ensure reproducibility is to comply with the NASA Open Science[8] guidelines.

In addition to the above recommendations, this work helps us understand the new research paths that the community can explore. It is important to keep in mind that this review presents the very first 24 ML models the community has developed to tackle the SEP prediction problem; therefore, it is not surprising that there are many unexplored research directions in such a nascent field. Based on the analysis presented in this document, a non-limiting list of open avenues for future exploration is given below:

- **Physics informed NNs:** Physics-Informed Neural Networks (PINNs) have proven to be a powerful approach to solving complex, non-linear scientific problems by embedding governing physical laws directly into the ML learning process (Raissi et al. 2019; Karniadakis et al. 2021). No studies that use PINNs in order to tackle the SEP prediction problem have been identified. Future studies should explore PINNs and other physics-based augmentations in ML models that predict SEP events.
- **Unsupervised models**: All existing work in the literature relies on labeled information (e.g., SEP lists or SEP onset/end times) to supervise model training. However, unsupervised learning offers a powerful alternative by removing the dependence on such externally provided targets, enabling models to discover intrinsic structure, latent representations, or precursor signatures directly from the data itself. Unsupervised ML techniques have already shown promise in heliophysics (Woods et al. 2021; Giger & Csillaghy 2024). Future work should explore SEP prediction using unsupervised approaches, which may reveal new physical insights and reduce biases introduced by manually curated event labels.
- **Inner heliosphere, Mars and cis-lunar environment predictions:** Twenty-two out of the twenty-four studies identified in this research aim to predict geoeffective SEP events, while only PSPSP predicts particle intensities across the heliosphere. With planned manned missions to the Moon and space-based infrastructure extending throughout the heliosphere, predicting SEP events away from Earth becomes increasingly important. Future studies should leverage the data we have from missions further away from the cis-lunar environment, such as the PSP, STEREO-A, and the Solar Orbiter (SolO; Müller et al. 2020; Marirrodriga et al. 2021), in order to predict SEP events in different parts of the heliosphere.
- **New observations:** A number of new heliophysics missions have been commissioned or planned. These missions will, or already do, provide us with new, higher quality data relevant to space weather. It is recommended that future SEP prediction works explore the usage of data from the most recent heliophysics missions such as the IMAP, PUNCH, SOLAR-1 (formerly called SWFO-L1) and Aditya-L1 (Tripathi et al. 2022) by the Indian Space Research Organisation (ISRO).
- **Real-time data:** Majority of the models discussed in this study are static; they have not been deployed for continuous real-time predictions. An example of real-time operational forecasting is provided by the SEPNET team through the University of Michigan Machine Learning for Space Weather (MLSW) website. The SEP forecasting model is run hourly using newly downloaded input features, and users can visualize the predictions on the website, along with the temporal trajectories of the input features and historical predictions. Future work should focus on deploying trained models in a similar fashion, utilizing on-demand, space missions that provide real-time observations, such as SDO, ACE, IMAP (I-ALiRT), STEREO-A, and others.
- **Pre-trained networks:** Most recently, the heliophysics community has trained large models, whose parameters and embeddings can be used as a base for developing new SEP prediction models. Future works could potentially use transfer learning approaches leveraging pre-trained heliophysics NNs or use the embeddings of Foundation Models (FMs) trained on heliophysics data (SuryaFM; Roy et al. 2025a,b) for predicting SEP events.

[8] https://science.nasa.gov/open-science/

- **Data augmentation methods:** SEPs are rare events and all models developed suffer from data imbalance. A rather unexplored area of research is the mitigation of the data imbalance problem in SEP prediction through data augmentation methods and production of synthetic but realistic SEP events which can be used for training ML models (see EPREM-S and TSF models in Appendix A.24 and A.16, respectively).
- **Connecting with other heliophysics predictions**: The applications of ML in heliophysics span from predicting solar flares (Jiao et al. 2020; Wang et al. 2020; Zheng et al. 2023), coronal mass ejections (Bobra & Ilonidis 2016; Vourlidas et al. 2019; Singh et al. 2023), the solar sunspot number (Sierra-Porta et al. 2024; Rodríguez et al. 2024; Qamar et al. 2025, SSN), the solar surface flux transport (Jeong et al. 2025), and even emerging ARs (Kasapis et al. 2023a, 2025a; Kosovichev et al. 2025; Tirona et al. 2026). These predictions are directly related to the production of SEP events. Future work should utilize such ML models to inform their SEP prediction efforts.

In summary, SEP prediction using ML remains an emerging field with promising but currently developing capabilities. Progress over the past decade demonstrates the community's growing ability to exploit heliophysical data using modern ML techniques. Continued advances will depend less on algorithmic advance and more on improvements in data integration, evaluation standardization and operational deployment. The insights presented here aim to guide future efforts towards reliable and operationally useful ML-based SEP forecasting systems.

6. CONCLUSIONS

This document reviews and summarizes (Appendix A) more than a decade of research that applies ML to the prediction of SEP events in the English-language literature. This community-wide effort includes descriptions of all identified models, along with tables that capture their quantitative performance and qualitative characteristics. The datasets used —or created specifically— for SEP prediction are also compiled and discussed (Section 3). Using this consolidated information, we provide a chartography of the current state of the SEP prediction using ML research community, mapping the landscape of model architectures, inputs, and outputs. We also attempt to compare these diverse approaches and their results, highlighting the difficulties and limitations encountered when making such comparisons. Based on this analysis, we outline recommended good practices for future studies and propose new research directions for the community to pursue.

This review compares three core aspects of existing SEP prediction studies —their inputs, the ML models trained on those inputs, and the outputs they produce. The community employs a wide range of architectures, largely shaped by available data and prediction goals (Section 4.1). By examining the inputs used (Section 4.2), we identify which missions and datasets have been most utilized, which observations are more accessible, and which potentially valuable data sources remain underused. Comparison of outputs (Section 4.3) illustrates a key challenge: the lack of standardized forecasting targets across studies, which makes direct comparison difficult. Sections 4.3 and 5 outline a framework to address this issue and suggest concrete directions for future research and progress toward operational ML-aided SEP forecasting.

To ensure that future work is comparable, reliable, and aligned with community and stakeholder's needs, we strongly recommend that new studies adopt the good practices outlined in this document, including the use of common validation metrics, proper uncertainty quantification, consistent prediction windows, larger and more modern ML architectures, operationally realistic validation setups, the use of SEPVAL for validation on common events, and full compliance with NASA Open Science principles. At the same time, our analysis reveals several promising research directions for the community to explore. These include the incorporation of PINNs, prediction efforts beyond geoeffective SEPs and into the broader heliosphere, the use of underutilized datasets and new mission observations, and leveraging pre-trained heliophysics FMs for transfer learning. Together, these recommendations and research avenues provide a clear path for advancing SEP prediction capabilities and strengthening the role of ML in heliophysics.

ACKNOWLEDGMENTS

This work originated from discussions held during the SEP Monitoring and Forecasting Workshop at Georgia State University, during October 16–19, 2024, for which Dr. Spiridon Kasapis received the AAS SPD Thomas Metcalf Travel Award Report, which we acknowledge along with the NASA AI/ML HECC Expansion Program, and the NASA grants 23-HGIO23_2-0077, 20-HSR20_2-0037, 80NSSC19K0630, 80NSSC19K0268, 80NSSC20K1870, and 80NSSC22M0162, which also supported Dr. Alexander Kosovichev and Dr. Irina Kitiashvili. This work was also supported, in part, by the IMAP mission as part of NASA's Solar Terrestrial Probes (STP) Program (80GSFC19C0027). Dr. Soukaina

Filali Boubrahimi, Shah Muhammad Hamdi, and Pouya Hosseinzadeh were supported in part by funding from the Division of Atmospheric and Geospace Sciences within the Directorate for Geosciences, under NSF awards #2204363, #2240022, #2530946, and #2301397. Dr. Angelos Vourlidas was supported by the NASA grants 80NSSC22K0970 and 80NSSC23K0412. Dr. Mohamed Nedal acknowledges support by the project "The Origin and Evolution of Solar Energetic Particles", funded by the European Office of Aerospace Research and Development under award No. FA8655-24-1-7392. Dr. Athanasios Papaioannou, Eleni Lavasa and Dr. Anastasios Anastasiadis received funding from the European Union's Horizon Europe programme under grant agreement No 101135044 (SPEARHEAD) [https://spearhead-he.eu/]. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Dr. Sumanth A. Rotti was supported by NASA FINESST grant No. 80NSSC21K1388 and SMD grant No. 24-SMDSS24-0045. Dr. Berkay Aydin and Dr. Petrus C. Martens were supported by NASA SWR2O2R grant No. 80NSSC22K0272. Dr. Christina Lee was supported in part by funding from NASA grant No. 80NSSC25K7646.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that this work complies with the ethical standards and policies outlined in the *Space Science Reviews* Instructions for Authors regarding compliance with ethical standards. The authors report no competing financial or non-financial interests that could have influenced the work presented in this manuscript. All co-authors have provided written consent to be included as authors and have approved the submitted version of the paper.

REFERENCES

Alberti, T., Laurenza, M., & Cliver, E. 2019, Nuovo Cimento C, 42, 40

Alberti, T., Laurenza, M., Cliver, E., et al. 2017, The Astrophysical Journal, 838, 59

Ali, A., Sadykov, V., Kosovichev, A., et al. 2024, The Astrophysical Journal Supplement Series, 270, 15

Ali, M. A., Abdelkawy, A. G., Shaltout, A. M., & Beheary, M. 2025, Scientific Reports, 15, 9546

Aminalragia-Giamini, S., Raptis, S., Anastasiadis, A., et al. 2021, Journal of Space Weather and Space Climate, 11, 59

Ash, A., & Shwartz, M. 1999, Statistics in medicine, 18, 375

Axford, W. I., Leer, E., & Skadron, G. 1977, in International Cosmic Ray Conference, Vol. 11, International Cosmic Ray Conference, 132

Bahri, O., Li, P., Hosseinzadeh, P., Boubrahimi, S. F., & Hamdi, S. M. 2023, in 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 453–458

Baydin, A. G., Poduval, B., & Schwadron, N. A. 2023, Space Weather, 21, e2023SW003593

Bell, A. R. 1978, MNRAS, 182, 147, doi: 10.1093/mnras/182.2.147

Bemporad, A. 2021, Frontiers in Astronomy and Space Sciences, 8, 627576

Benella, S., Stumpo, M., Laurenza, M., et al. 2023, Proceedings of Science (ECRS)

Benesty, J., Chen, J., Huang, Y., & Cohen, I. 2009, in Noise reduction in speech processing (Springer), 1–4

Berger, T., Camporeale, E., Poduval, B., Delouille, V. A., & Murray, S. A. 2021, Machine Learning in Heliophysics (Frontiers Media SA)

Blandford, R. D., & Ostriker, J. P. 1978, ApJL, 221, L29, doi: 10.1086/182658

Bobra, M. G., & Ilonidis, S. 2016, The Astrophysical Journal, 821, 127

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, Solar Physics, 289, 3549

Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. 2021, The Astrophysical Journal Supplement Series, 256, 26

Boubrahimi, S. F., Aydin, B., Martens, P., & Angryk, R. 2017, in 2017 IEEE international conference on big data (big data), IEEE, 2533–2542

Bougeret, J.-L., Kaiser, M. L., Kellogg, P. J., et al. 1995, Space Science Reviews, 71, 231

Brueckner, G., Howard, R., Koomen, M., et al. 1995, Solar Physics, 162, 357

Bruno, A. 2017, Space Weather, 15, 1191, doi: 10.1002/2017SW001672

Burt, J., & Smith, B. 2012, in 2012 ieee aerospace conference, IEEE, 1–13

Buzulukova, N., & Tsurutani, B. 2022, Frontiers in Astronomy and Space Sciences, 9, 1017103

Cabello, N., Naghizade, E., Qi, J., & Kulik, L. 2020, in 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 948–953

Camporeale, E., & of ML-Helio, S. O. C. 2020, Journal of Geophysical Research: Space Physics, 125, e2019JA027502

Chamberlin, P. C., Woods, T. N., Eparvier, F. G., & Jones, A. R. 2009, in Solar physics and space weather instrumentation III, Vol. 7438, SPIE, 11–20

Chatterjee, S., Dayeh, M. A., Muñoz-Jaramillo, A., et al. 2024, Space Weather, 22, e2023SW003568

Chhiber, R., Matthaeus, W. H., Cohen, C. M. S., et al. 2021, A&A, 650, A26, doi: 10.1051/0004-6361/202039816

Cho, K.-S., Hwang, J., Han, J.-Y., et al. 2023, Journal of the Korean Astronomical Society, 56, 263

Cohen, C., Alterman, B. L., Baker, D. N., et al. 2026, Space Science Reviews, 222, 6

Cohen, C. M. S., Christian, E. R., Cummings, A. C., et al. 2021a, A&A, 650, A23, doi: 10.1051/0004-6361/202039299

—. 2021b, A&A, 656, A29, doi: 10.1051/0004-6361/202140967

Creech, S., Guidi, J., & Elburn, D. 2022, in 2022 ieee aerospace conference (aero), IEEE, 1–7

Cucinotta, F. A., Kim, M.-H. Y., Chappell, L. J., & Huff, J. L. 2013, PloS one, 8, e74988

Cuesta, M., Khoo, L., Livadiotis, G., et al. 2025, The Astrophysical Journal, 980, 235

Cuesta, M. E., Fraschetti, F., Livadiotis, G., et al. 2025, ApJL, 993, L15, doi: 10.3847/2041-8213/ae109c

Dayeh, M. A., Chatterjee, S., Muñoz-Jaramillo, A., et al. 2024, Space Weather, 22, e2023SW003697

Dayeh, M. A., Starkey, M. J., Chatterjee, S., et al. 2025, arXiv preprint arXiv:2502.08555

DeForest, C., Killough, R., Gibson, S., et al. 2022, in 2022 IEEE Aerospace Conference (AERO), IEEE, 1–11

Delaboudiniere, J.-P., Artzner, G., Brunaud, J., et al. 1995, Solar Physics, 162, 291

Desai, M., & Giacalone, J. 2016, Living Reviews in Solar Physics, 13, 3

Domingo, V., Fleck, B., & Poland, A. 1995a, Space Science Reviews, 72, 81

Domingo, V., Fleck, B., & Poland, A. I. 1995b, Solar Physics, 162, 1

Doswell, C., Davies-Jones, R., & Keller, D. L. 1990, Weather and forecasting, 5, 576

Drury, L. O. 1983, Reports on Progress in Physics, 46, 973

Engelbrecht, N. E., Effenberger, F., Florinski, V., et al. 2022, Space Science Reviews, 218, 33

Engell, A., Falconer, D., Schuh, M., Loomis, J., & Bissett, D. 2017, Space Weather, 15, 1321

Forrester, A., Sobester, A., & Keane, A. 2008, Engineering design via surrogate modelling: a practical guide (John Wiley & Sons)

Fox, N., Velli, M., Bale, S., et al. 2016, Space Science Reviews, 204, 7

Georgoulis, M. K., Yardley, S. L., Guerra, J. A., et al. 2024, Advances in Space Research

Giger, M., & Csillaghy, A. 2024, Space Weather, 22, e2023SW003516

Goel, M. K., Khanna, P., & Kishore, J. 2010, International journal of Ayurveda research, 1, 274

Gopalswamy, N. 2022, Atmosphere, 13, 1781

Grim, L. F. L., & Gradvohl, A. L. S. 2024, Solar Physics, 299, 33

Hanser, F. 2011, Tech. Rep. GOESN-ENG-048D

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778

Hill, M. E., Mitchell, D. G., Andrews, G. B., et al. 2017, Journal of Geophysical Research (Space Physics), 122, 1513, doi: 10.1002/2016JA022614

Hodson, T. O. 2022, Geoscientific Model Development Discussions, 2022, 1

Hosseinzadeh, P., Boubrahimi, S. F., & Hamdi, S. M. 2024a, The Astrophysical Journal Supplement Series, 270, 31

—. 2025, The Astrophysical Journal Supplement Series, 277, 34

Hosseinzadeh, P., Filali Boubrahimi, S., & Hamdi, S. M. 2024b, Space Weather, 22, e2024SW003982

Hu, S., & Semones, E. 2022, Journal of Space Weather and Space Climate, 12, 5

Hutchins, T., & Kasapis, S. 2026, Solar Energetic Particle Prediction Using Deep Learning and PSP/ISOIS Data, V1, AGU, doi: 10.0000/XXX/XXXXXX

Jackson, I., & Martens, P. 2024a, The Astrophysical Journal Supplement Series, 272, 37

—. 2024b, Survival Solar Energetic Particle (SSEP) Dataset, V1, Harvard Dataverse, doi: 10.7910/DVN/GXY9MZ

Jeong, H.-J., Jeon, M., Kim, D., et al. 2025, The Astrophysical Journal Supplement Series, 278, 5

Jiao, Z., Sun, H., Wang, X., et al. 2020, Space weather, 18, e2020SW002440

JOSELYN, J., & GRUBB, R. 1985, in 23rd Aerospace Sciences Meeting, 238

Kaiser, M. L., Kucera, T., Davila, J., et al. 2008, Space Science Reviews, 136, 5

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., et al. 2021, Nature Reviews Physics, 3, 422

Kasapis, S., Kitiashvili, I. N., Kosovich, P., et al. 2024, The Astrophysical Journal, 974, 131

Kasapis, S., Kitiashvili, I. N., Kosovichev, A. G., & Stefan, J. T. 2025a, The Astrophysical Journal Supplement Series, 280, 64

Kasapis, S., Kitiashvili, I. N., Kosovichev, A. G., Stefan, J. T., & Apte, B. 2023a, Proceedings of the International Astronomical Union, 19, 311

Kasapis, S., Thompson, B. J., Rodriguez, J. V., et al. 2023b, Space Weather, 21, e2022SW003310

Kasapis, S., Zhao, L., Chen, Y., et al. 2022, Space Weather, 20, e2021SW002842

Kasapis, S., Cuesta, M. E., Khoo, L. Y., et al. 2025b, The Astrophysical Journal

Kasper, J. C., Abiad, R., Austin, G., et al. 2016, Space Science Reviews, 204, 131

Keegan, K., Bonaventura, N., Guzmán, P., et al. 2025, in NeurIPS 2025 AI for Science Workshop

Kosovich, P. A., Kosovichev, A. G., Sadykov, V. M., et al. 2024, The Astrophysical Journal, 972, 169

Kosovichev, A. G., Basu, S., Bekki, Y., et al. 2025, Solar Physics, 300, 70

Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, Advances in neural information processing systems, 30

Laurenza, M., Alberti, T., & Cliver, E. 2018, The Astrophysical Journal, 857, 107

Laurenza, M., Cliver, E., Hewitt, J., et al. 2009, Space Weather, 7

Laurenza, M., Stumpo, M., Zucca, P., et al. 2024, Journal of Space Weather and Space Climate, 14, 8

Lavasa, E., Giannopoulos, G., Papaioannou, A., et al. 2021, Solar Physics, 296, 107

Lee, C. O., Christian, E. R., Sandoval, L., et al. 2025, Space Science Reviews, 221, 117, doi: 10.1007/s11214-025-01244-9

Lemen, J. R., Title, A. M., Akin, D. J., et al. 2012, Solar Physics, 275, 17

Li, P., Bahri, O., Boubrahimi, S. F., & Hamdi, S. M. 2025, The Astrophysical Journal Supplement Series, 280, 52

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. 2014, stat, 1050, 14

Livi, R., Larson, D. E., Kasper, J. C., et al. 2022, The Astrophysical Journal, 938, 138

Löning, M., Bagnall, A., Ganesh, S., et al. 2019, arXiv preprint arXiv:1909.07872

Löning, M., Király, F., Bagnall, T., et al. 2022, sktime/sktime: v0.13.4, Zenodo, doi: 10.5281/zenodo.7117735

Macmillan, N. A., & Kaplan, H. L. 1985, Psychological bulletin, 98, 185

Malandraki, O. E., & Crosby, N. B. 2018, Solar particle radiation storms forecasting and analysis: The HESPERIA HORIZON 2020 project and beyond (Springer Nature)

Marirrodriga, C. G., Pacros, A., Strandmoe, S., et al. 2021, Astronomy & Astrophysics, 646, A121

McComas, D., Alexander, N., Angold, N., et al. 2016, Space Science Reviews, 204, 187

McComas, D., Christian, E. R., Schwadron, N. A., et al. 2018, Space science reviews, 214, 116

McComas, D. J., Christian, E. R., Cohen, C. M. S., et al. 2019, Nature, 576, 223, doi: 10.1038/s41586-019-1811-1

McComas, D. J., Christian, E. R., Schwadron, N., et al. 2025, Space science reviews, 221, 100

Meyer, G. P. 2021, in Proceedings of the ieee/cvf conference on computer vision and pattern recognition, 5261–5269

Miroshnichenko, L. I. 2018, Journal of Space Weather and Space Climate, 8, A52

Mishev, A., Adibpour, F., Usoskin, I., & Felsberger, E. 2015, Advances in Space Research, 55, 354

Mitchell, J. G., Christian, E. R., de Nolfo, G. A., et al. 2025, ApJ, 980, 96, doi: 10.3847/1538-4357/adaa7c

Moreland, K., Dayeh, M. A., Bain, H. M., et al. 2024, Space Weather, 22, e2023SW003765

Muschelli III, J. 2020, Journal of classification, 37, 696

Müller, D., St. Cyr, O. C., Zouganelis, I., et al. 2020, Astronomy & Astrophysics, 642, A1, doi: 10.1051/0004-6361/202038467

Narock, A., Bard, C., Thompson, B. J., et al. 2022, Frontiers in Astronomy and Space Sciences, 9, 1064233

Nedal, M., Kozarev, K., Arsenov, N., & Zhang, P. 2023, Journal of Space Weather and Space Climate, 13, 26

Neukart, F. 2024, Heliyon, 10

Nita, G., Georgoulis, M., Kitiashvili, I., et al. 2020, arXiv preprint arXiv:2006.12224

Núñez, M. 2011, Space Weather, 9

Núñez, M., & Paul-Pena, D. 2020, Universe, 6, 161

O'Keefe, P. M., Sadykov, V., Kosovichev, A., et al. 2024, Advances in Space Research, 74, 6252

Pacheco Mateo, D. 2019

Pak, S., Cuesta, M., Farooki, H., et al. 2025, The Astrophysical Journal Supplement Series, 281, 21

Panigrahi, R., & Borah, S. 2018, Procedia computer science, 132, 323

Papaioannou, A., Strauss, R. D. T., Lario, D., et al. 2025, SSRv, 221, 82, doi: 10.1007/s11214-025-01211-4

Papaioannou, A., Sandberg, I., Anastasiadis, A., et al. 2016, Journal of Space Weather and Space Climate, 6, A42

Peirce, C. S. 1884, Science, 4, 453, doi: 10.1126/science.ns-4.93.453

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. 2012, in The solar dynamics observatory (Springer), 3–15
Posner, A., Arge, C. N., Staub, J., et al. 2021, Space Weather, 19, e2021SW002777
Qamar, W., Hussain, M., Zaheer, M. B., et al. 2025, Astrophysics and Space Science, 370, 68
Queipo, N. V., Haftka, R. T., Shyy, W., et al. 2005, Progress in aerospace sciences, 41, 1
Raissi, M., Perdikaris, P., & Karniadakis, G. E. 2019, Journal of Computational physics, 378, 686
Rankin, J. S., Bindi, V., Bykov, A. M., et al. 2022, Space Science Reviews, 218, 42
Raouafi, N. E., Matteini, L., Squire, J., et al. 2023, Space Science Reviews, 219, 8
Reames, D. V. 2013, Space Science Reviews, 175, 53
—. 2021, Solar energetic particles: a modern primer on understanding sources, acceleration and propagation (Springer Nature)
Regnault, F., Janvier, M., Démoulin, P., et al. 2020, Journal of Geophysical Research: Space Physics, 125, e2020JA028150
Richardson, I., Mays, M., & Thompson, B. 2018, Space Weather, 16, 1862
Richardson, I., Von Rosenvinge, T., Cane, H., et al. 2014, Solar Physics, 289, 3059
Rodriguez, J., Onsager, T., & Mazur, J. 2010, Geophysical Research Letters, 37
Rodríguez, J.-V., Sánchez Carrasco, V. M., Rodríguez-Rodríguez, I., Pérez Aparicio, A. J., & Vaquero, J. M. 2024, Solar Physics, 299, 116
Rotti, S., Aydin, B., Georgoulis, M., & Martens, P. 2022a, GSEP Dataset, V5, Harvard Dataverse, doi: 10.7910/DVN/DZYLHK
Rotti, S., Aydin, B., Georgoulis, M. K., & Martens, P. C. 2022b, The Astrophysical Journal Supplement Series, 262, 29
Rotti, S., & Martens, P. C. 2023, The Astrophysical Journal Supplement Series, 267, 40
Rotti, S. A., Aydin, B., & Martens, P. C. 2024a, The Astrophysical Journal, 966, 165
—. 2024b, The Astrophysical Journal, 974, 188
Roy, S., Schmude, J., Lal, R., et al. 2025a, arXiv preprint arXiv:2508.14112
Roy, S., Hegde, D. V., Schmude, J., et al. 2025b, arXiv preprint arXiv:2508.14107
Ryan, D. F., Milligan, R. O., Gallagher, P. T., et al. 2012, The Astrophysical Journal Supplement Series, 202, 11
—. 2013, SDO-3: Exploring the Network of SDO Science, 143
Sadykov, V., Kosovichev, A., Kitiashvili, I., et al. 2021, arXiv preprint arXiv:2107.03911
Sadykov, V. M., Kosovichev, A. G., Kitiashvili, I. N., & Frolov, A. 2019, The Astrophysical Journal, 874, 19
Sandberg, I., Jiggens, P., Heynderickx, D., & Daglis, I. A. 2014, Geophys. Res. Lett., 41, 4435, doi: 10.1002/2014GL060469
Scherrer, P. H., Bogart, R. S., Bush, R., et al. 1995, Solar Physics, 162, 129
Scherrer, P. H., Schou, J., Bush, R., et al. 2012, Solar Physics, 275, 207
Schrijver, C. J., Kauristie, K., Aylward, A. D., et al. 2015, Advances in Space Research, 55, 2745
Schwadron, N. A., Townsend, L., Kozarev, K., et al. 2010, Space Weather, 8
Sellers, F. B., & Hanser, F. A. 1996, in GOES-8 and Beyond, Vol. 2812, SPIE, 353–364
Sierra-Porta, D., Tarazona-Alvarado, M., & Acevedo, D. H. 2024, Astronomy and Computing, 48, 100857
Simunac, K., & Armstrong, T. 2004, Journal of Geophysical Research: Space Physics, 109
Singh, T., Benson, B., Raza, S. A., et al. 2023, The Astrophysical Journal, 948, 78
Stone, E. C., Frandsen, A., Mewaldt, R., et al. 1998, Space Science Reviews, 86, 1
Stumpo, M., Benella, S., Laurenza, M., et al. 2021, Space Weather, 19, e2021SW002794
Subashchandar, N. S. M., Zhao, L., Shalchi, A., et al. 2025, ApJL, 991, L30, doi: 10.3847/2041-8213/ae063f
Temmer, M. 2021, Living Reviews in Solar Physics, 18, 4
Tirona, J., Patil, S., Kasapis, S., et al. 2026, arXiv preprint arXiv:2601.13144
Tobiska, W. K., Atwell, W., Beck, P., et al. 2015, Space Weather, 13, 202
Torres, J., Chan, P. K., Zhao, L., & Zhang, M. 2025, Space Weather, 23, e2024SW003921
Torres, J., Zhao, L., Chan, P. K., & Zhang, M. 2022, Space Weather, 20, e2021SW002797
Tripathi, D., Chakrabarty, D., Nandi, A., et al. 2022, Proceedings of the International Astronomical Union, 18, 17
van Haarlem, M. P., Wise, M. W., Gunst, A., et al. 2013, Astronomy & astrophysics, 556, A2
Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, Advances in neural information processing systems, 30
Von Rosenvinge, T., Barbier, L., Karsch, J., et al. 1995, Space Science Reviews, 71, 155
Vourlidas, A., Patsourakos, S., & Savani, N. 2019, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 377

Wang, X., Chen, Y., Toth, G., et al. 2020, The Astrophysical Journal, 895, 3

Waterfall, C. O. G., Dalla, S., Raukunen, O., et al. 2023, Space Weather, 21, e2022SW003334, doi: 10.1029/2022SW003334

Wehling, P., LaBudde, R. A., Brunelle, S. L., & Nelson, M. T. 2011, Journal of AOAC International, 94, 335

Wen, J., & Angryk, R. A. 2024, in International Conference on Artificial Intelligence and Soft Computing, Springer, 362–375

Whitman, K., Egeland, R., Allison, C., Quinn, P., & Stegeman, L. 2026, NASA Technical Reports Server

Whitman, K., Egeland, R., Richardson, I. G., et al. 2023, Advances in Space Research, 72, 5161

Woods, M. M., Sainz Dalda, A., & De Pontieu, B. 2021, The Astrophysical Journal, 922, 137

Woods, T. N., Eden, T., Eparvier, F. G., et al. 2024, Journal of Geophysical Research: Space Physics, 129, e2024JA032925

Yu, Y., Chen, Y., Zhao, L., et al. 2025, arXiv preprint arXiv:2512.12786

Zank, G., Hunana, P., Mostafavi, P., et al. 2015, The Astrophysical Journal, 814, 137

Zeitlin, C., Hassler, D., Cucinotta, F., et al. 2013, science, 340, 1080

Zheng, Y., Qin, W., Li, X., et al. 2023, Astrophysics and Space Science, 368, 53

APPENDIX

A. DESCRIPTION OF ML MODELS

The descriptions of the models presented here are based on the modelers' contributions and their answers to the questionnaire presented in Appendix C. The models here appear in order of complexity, from less complex to deeper, as they are summarized in Table 1. Each of the following subsections (Sections A.1-A.24) presents a single-page summary of the ML models that predict SEP events, along with a table of quantitative and qualitative characteristics of each models, as summarized in Table 3 and Figures 2 and 3.

A.1. *eXtreme Gradient Boosting (XGBoost) Model*

Model Developers and Relevant Citation: Aatiya Ali, Viacheslav Sadykov, Alexander Kosovichev, Irina N. Kitiashvili, Vincent Oria, Gelu M. Nita, Egor Illarionov, Patrick M. O'Keefe, Fraila Francis, Chun-Jie Chong, Paul Kosovich, and Russell D. Marroquin; Ali et al. (2024).

Table 6: Model, Input and Output Specification Table for the XGBoost model.

Model	
Type	Decision Tree
Complexity	2
Input	
Shape	Time Series (1D)
Type	Soft X-ray, Proton Flux
History	33 years (1986-2019)
Diversity	12484 samples
Imbalance	0.045 positive
Sample Size	240 bytes
Sample Coverage	24 hours
Output	
Prediction	Classification
Type	Continuous
Forecast Window	23 hours

Summary: The eXtreme Gradient Boosting (XGBoost; Figure 6a) model generates binary predictions of SPEs for the following day at Geostationary Earth Orbit (GEO). Comparing the performance of an SVM and XGBoost for these predictions, we find that XGBoost significantly outperforms SVM in most training-testing configurations based on metrics such as TSS, HSS, and recall. Using GOES proton and soft X-ray flux data spanning from 1986 to 2019. Data from SCs 22-24 are treated separately for training and testing. To address class imbalance, various oversampling and weight-balancing methods were tested, as well as model cross-cycle transferability.

Model Description: XGBoost —an ensemble classifier based on gradient boosting, presents better results compared to SVMs (supervised classifiers using decision surfaces), across all tests: using default parameters, applying imbalance-handling class weights, and using data oversampled by standard (positive-class duplication), ADAptive SYNthetic (ADASYN), and Synthetic Minority Oversampling TEchniques (SMOTE) separately. Flux feature importance is determined using Gini importance, Fisher scoring, and the inherent feature ranking provided by XGBoost. Because our primary goal is not to parameterize a specific algorithm with minute detail, only parameters related to classification are optimized using Grid Search Cross-Validation (*GridSearchCV*[9]).

Inputs: Model input features include statistics of (1-8 Å) soft X-ray fluxes and proton fluxes $\geq$ 10 MeV from the Energetic Particle Sensors (EPS; Hanser 2011) and Energetic Proton, Electron and Alpha Detectors (EPEAD; Bruno

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

2017) onboard the GOES missions. These features include daily flux mean, median, minimum, maximum, standard deviation, skewness, kurtosis, and the last measured flux of the previous day. The dataset exhibits a significant class imbalance, with 11,946 days classified as negative (no SPEs) and only 538 days classified as positive (SPEs detected).

Outputs: The model produces a daily binary flag, indicating whether an SPE is expected at GEO in the next 23 hours.

Model Configuration: The XGBoost model specifies two default parameters: booster = gbtree (to use tree-based models for ensemble building) and scale_pos_weight (to balance data classes).

Model Validation and Results: On average, XGBoost outperformed SVM by approximately +0.10 in TSS, +0.20 in HSS, and +0.10 in recall. While XGBoost showed higher recall values, it also produced a significant number of false positives. Evaluation of the XGBoost model across long (two SCs) and short (a single SC) training timescales shows that TSS and HSS were comparable for both timescales. The cross-cycle transferability studies the dependence of the results on properties of a solar cycle. Nonetheless, when compared to baseline models, such as SWPC daily probabilistic forecasts and a persistence model, XGBoost (optimized for TSS) outperformed these models in both TSS and recall. Overall, our results suggest that with proper tuning, XGBoost can enhance SPE prediction accuracy, particularly in refining all-clear predictions.

Access to Model Data and Forecasts: The SPE catalogs developed during this study are archived at: https://sun.njit.edu/SEP3/datasets.html. The GOES proton and soft X-ray flux data are available at http://www.ncei.noaa.gov/data/goes-space-environment-monitor/access/avg/. The XGBoost python package is accessible via https://xgboost.readthedocs.io/en/stable/python/python_intro.html.

Limitations, Caveats and Discussion: This work builds on NOAA's classification of an S1 SPE where protons $\geq$ 10 MeV exceed 10 pfu. The model results were tested and validated with GOES flux data from only SCs 22-24 individually across different training and testing configurations. Further work is needed to improve the model's ability to generalize across multiple SCs with varying levels of solar activity.

A.2. *Supervised Time Series Forest (STSF) Model*

Model Developers and Relevant Citation: Sumanth A. Rotti, Berkay Aydin, and Petrus C. Martens; Rotti et al. (2024a,b) and Löning et al. (2019).

Table 7: Model, Input and Output Specification Table for the STSF model.

Model	
Type	Forest
Complexity	4
Input	
Shape	Time Series (1D)
Type	X-ray, Energetic Protons
History	32 years (1986-2018)
Diversity	998 samples
Imbalance	0.09 positive
Sample Size	25000 bytes
Sample Coverage	11 hours
Output	
Prediction	Classification
Type	Continuous
Forecast Window	1 hour

Summary: In Rotti et al. (2024a), an ensemble modeling methodology is introduced, consisting of a feature-based multivariate variant of univariate time series classifiers to classify between strong and weak SEP events in the GSEP data set (Rotti et al. 2022a,b; Rotti & Martens 2023) covering SCs 22-24. The same model architecture was implemented in Rotti et al. (2024b) on an extended dataset comprising SEP-quiet samples. There are 2893 samples, of which 244 are strong SEP events (those crossing the SWPC's S1 threshold). Furthermore, a fixed input length of

the time series (11 hours) is considered and the model is assessed for a prediction window of up to 60 minutes. The model's performance on an expanded dataset is promising, obtaining high skill scores.

Model Description: Both studies utilized a binary classification framework (SEPs vs. weak/non-SEPs) using an ensemble of univariate time series classifiers. The best-performing model is the Supervised Time Series Forest (STSF Cabello et al. 2020), which is described in this review. The STSF model employs three representations (time, frequency, and derivative) of the input time series and uses a supervised learning approach to find discriminatory intervals. The model computes the region of interest to highlight the location of discriminatory intervals, defined as the intersection of such intervals. Furthermore, it extracts seven statistical features, including mean, median, standard deviation, slope, minimum, maximum, and interquartile range, from each interval. The ranking of the interval feature is determined by a scoring function that indicates how effectively the feature distinguishes one class of time series from the other classes. The final set of intervals is obtained in a top-down approach to represent the entire series. The feature set is concatenated to form a new dataset upon which decision trees are built. The final output is based on the majority vote of averaged probability estimates from the individual estimators in the ensemble.

Inputs: In our approach, strong SEPs (244 samples) correspond to the positive class, while weak and non-SEPs (2,649 samples) are negatives. Here, a strong SEP-event indicates the GOES $\geq$ 10 MeV proton fluxes crossing 10 pfu. The data set contains long-band (1-8Å) X-ray measurements from the XRS instrument and proton fluxes from the Space Environment Monitor (SEM; JOSELYN & GRUBB 1985) instrument onboard the GOES missions. The model is trained using these four physical parameters as input.

Outputs: The output of the ensemble STSF model is a binary flag (yes/no SEP). That is, it indicates whether a strong SEP event will occur within the next 60 minutes, based on the observed X-ray and proton enhancements.

Model Configuration: The STSF model is available in the `sktime` library (Löning et al. 2019; Löning et al. 2022) for Python. We use the training set to train the model and perform a grid search for hyperparameter optimization. The best hyperparameters for STSF were in the default model settings, with the number of estimators set to 200.

Model Validation and Results: The model utilizes TSS, HSS, Gilbert Skill Score (GSS), and Matthew's Correlation Coefficient (MCC) scores for evaluation. For a 60-minute prediction window, the scores are TSS = 0.850, HSS = 0.878, GSS = 0.783, and MCC = 0.879. This study examines periods of non-occurrence of SEPs following a flare with magnitudes $\geq C6.0$ to maintain a natural class imbalance in the sample distribution. Nonetheless, there was only a decrease of ~ 7 % ($\pm 2\%$) in the skill scores compared to Rotti et al. (2024a).

Access to Model Data and Forecasts: The GSEP data set (Rotti et al. 2022a) and coding methodology of our model implementation have been made available on the GitHub repository: https://github.com/sumanth-ra23/SEP-Predictions. The SEP catalog and time series data set developed as a part of this study are available on Harvard Dataverse: https://doi.org/10.7910/DVN/DZYLHK.

Limitations, Caveats and Discussion: Identifying and flagging strong SEP events with proton fluxes fluctuating around 10 pfu is challenging for our model, which can lead to some misses and false positives.

A.3. *SHARP-SMARP Model*

Model Developers and Relevant Citation: Spiridon Kasapis, Lulu Zhao, Yang Chen, Xiantong Wang, Monica Bobra, Tamas Gombosi, Irina N. Kitiashvili, Paul Kosovich, Alexander G. Kosovichev, Viacheslav M. Sadykov, Patrick O'Keefe, and Vincent Wang; Kasapis et al. (2022, 2024) and Kosovich et al. (2024).

Summary: The 2022 model (Kasapis et al. 2022) introduced an interpretable ML framework using the SMARP (SDO/MDI) dataset to forecast whether solar flares would lead to SEP events during SC 23. The 2024 model (Kasapis et al. 2024) extended this work to include SHARP (SDO/HMI) data using the Kosovich et al. (2024) dataset, allowing prediction of SEPs across SCs 23 and 24 with a combined dataset of 3,869 ARs and 110 SEP events, twice as many as the 2022 study. Despite the expanded dataset, the model performance remained similar, indicating a limit to the predictive power of the selected SHARP and SMARP features.

Model Description: Both studies (Kasapis et al. 2022, 2024) utilized a binary classification framework (SEP and non-SEP warnings after a flare occurrence) by using SVMs and Regression models. Different kernels were explored for the SVMs such as a) the linear kernel, b) the polynomial kernel, c) the radial basis function, and d) the sigmoid kernel, while both logistic and linear regression models were also tested. The best performing model mentioned in this review is the linear SVM.

Inputs: The approach defines flares associated with SEPs as positive cases (110 positive flares), and flare-only events (3356 negative flares) as negatives. SMARP and SHARP Bobra et al. (2021) physical parameters are used

Table 8: Model, Input and Output Specification Table for the SHARP-SMARP model.

Model	
Type	Support Vector Machine
Complexity	7
Input	
Shape	Point Data (0D)
Type	Magnetic Fields
History	26 years (1996-2022)
Diversity	3466 samples
Imbalance	0.032 positive
Sample Size	56 bytes
Sample Coverage	0 hours
Output	
Prediction	Classification, Probability
Type	Triggered
Forecast Window	14.21 hours
Comments: *The forecast window is variable, defined by the model's prediction timestamp until the beginning of the SEP event. The average forecast window is 14.21 hours.	

as predictive features, spanning 26 years (1996-2022). The model was trained using five physical features from the SMARP-SHARP dataset: the total line-of-sight unsigned flux, the mean value of line-of-sight magnetic field gradient, the unsigned flux R near polarity inversion lines, the Vertical component of the total unsigned magnetic flux, and the mean value of the vertical field gradient. Two more parameters are calculated using the AR coordinates, the AR area, and the AR's angular distance between the associated flare and Earth's magnetic footpoint. From the full SMARP-SHARP timelines for the aforementioned physical parameters provided by Kosovich et al. (2024), only those recorded right before each flare occurrence were selected. Therefore, for each positive or negative flare instance, a 7-dimensional vector (56 bytes) is used for training and testing the SVM model.

Outputs: The output of the model is a probability of SEP occurrence based on a flare trigger. The probability is converted to a binary label (True/False) using a threshold of 0.5, informing us whether a flare will produce an SEP event.

Model Configuration: The model that yielded the best results in our analysis is an SVM that uses a linear kernel. For the 7-dimensional data case, the SVM uses 7 trainable parameters (weights). The Python Scikit Learn (*sklearn*[10]) library was used for implementation along with the *GridSearchCV* in order to explore the best regularization parameter C=2.4 using the standard l2 penalty.

Model Validation and Results: The study shows that despite the augmented volume of data compared to Kasapis et al. (2022), the prediction accuracy reaches 0.7 ± 0.1 (experimental setting / balanced dataset), which aligns with but does not exceed these published benchmarks. A linear SVM model with training and testing configurations that mimic an operational setting (original positive–negative imbalance) reveals a slight increase (0.04 ± 0.05) in the accuracy of a 14-hour SEP forecast compared to Kasapis et al. (2022). Other metrics used in the study are: TSS, HSS, FAR and F1.

Access to Model Data and Forecasts: Data and preprocessing scripts are available at github.com/skasapis/SEP_Pred_SMARP-SHARP. The SEP list used in this study is provided by NOAA at https:/ngdc.noaa.gov/stp/satellite/goes/doc/SPE.txt while the SMARP-SHARP dataset, although not publicly available yet, can be obtained by contacting the authors of Kosovich et al. (2024).

Limitations, Caveats and Discussion: A major limitation of both the SMARP-SHARP models is the inherent class imbalance in SEP prediction: only a small fraction of flaring ARs actually produce SEPs. This imbalance can bias the models toward predicting non-events unless careful sampling or penalization strategies are employed. While

[10] https://scikit-learn.org/stable/

physical parameters like magnetic flux and magnetic field gradients show some predictive capability, they are not sufficient to achieve high skill scores, even when expanding the dataset from one to two SCs. The plateau in model performance between the two studies suggests a ceiling on what can be achieved with current inputs and models alone.

A.4. *AA Model*

Model Developers and Relevant Citation: Eleni Lavasa, Giorgos Giannopoulos, Athanasios Papaioannou, Anastasios Anastasiadis; Lavasa et al. (2021).

Table 9: Model, Input and Output Specification Table for the AA model.

Model	
Type	Forest*
Complexity	8
Input	
Shape	Point data (0D)
Type	Soft X-ray, Coronagraphs
History	15 years (1998–2013)
Diversity	3,307 samples
Imbalance	0.039 positive
Sample Size	64 bytes
Sample Coverage	0 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	24 hours
Comments: *The sklearn.ensemble.RandomForestClassifier class from the *sklearn* library was used here.	

Summary: The Random Forest (RF) model generates binary predictions of solar energetic particle (SEP) events (integral proton flux ≥ 10 pfu at $E \geq 10$ MeV), for the following day at GEO. Comparing the performance of several ML algorithms (logistic regression, SVM, decision tree, random forest, extremely randomized trees, XGBoost and NNs), we find that random forests show both high performance in terms of F1-score, TSS, HSS (achieving high POD with relatively low FAR) and generalization (i.e. low variance), in the mean test scores of a 5-fold nested cross-validation scheme. Random forests also show a small difference in performance between the validation and test sets. The model is trained, validated and tested on flare soft X-ray data from GOES and CME data from SOHO/LASCO spanning from 1998 to 2013 (SC 23 and the rising phase of SC 24). Weight-adjusting to more heavily penalize misclassified positive SEP events is applied to address class imbalance.

Model Description: Random Forests —a bagging ensemble classifier built on decision trees— are the method of choice among other supervised classifiers (logistic regression, SVM, decision tree, extremely randomized trees, XGBoost and NNs) in this evaluation setting. A nested 5-fold cross-validation scheme is used, with inner folds for hyperparameter tuning (under *RandomizedGridSearchCV*[11] across 1000 configurations) and outer folds for the evaluation of model performance and generalization. Data are split randomly without replacement to prevent information leakage and stratified to preserve class imbalance across all partitions. Permutation feature importance, measuring the decrease in model performance when a feature's values are randomly shuffled, is applied to assess the importance of input features to the prediction target.

Inputs: Solar eruptive flare and CME events associated with SEPs are defined as positive cases (1 label) and eruptive events without SEP association as negatives (0 label). Model input features are extracted from solar flare identifications (classes C, M, X) in the 1-8 Åsoft X-ray flux from the XRS instrument[12] onboard the GOES missions, complemented with locations extracted in Ha, as well as CME recordings in white light by the SOHO/LASCO coronagraph. These

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

[12] ftp://ftp.ngdc.noaa.gov/STP/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/

features include the peak flux, fluence (time-integrated flux), heliographic longitude, duration, and the rise time of solar flares, as well as the sky-projected linear speed and width of the CME. An additional Cycle index is used as indicative to the magnitude of solar activity. The dataset includes a significant class imbalance (∼3.9%), with 3181 negative (no SEP) and only 126 positive (SEP) class events, thus, being representative of the real distribution of events.

Outputs: Given parent solar event triggers ($\geq C1$ flare, CME), the model produces a binary flag, indicating whether an SEP event is expected at Earth in the next 24 hours, exceeding integral proton flux ≥ 10 pfu at $E \geq 10$ MeV.

Model Configuration: The following hyper-parameters of the random forest classifier are optimized for F1-score t: i) n_estimators = total number of decision trees in the ensemble, ii) criterion = split criterion in individual decision trees, iii) min_samples_split = minimum number of samples required to perform a split, iv) min_samples_leaf = min. number of samples in decision tree leaf nodes, v) max_depth = maximum depth of decision trees, vi) class_weight controls the strength of penalization applied to wrong predictions of positive and negative class events, vii) max_features = maximum number of features in constructing individual decision trees.

Model Validation and Results: On the realistic, imbalanced test folds, random forests emerged as the top-scoring classifier. Using the combined flare and CME feature set without imputed gaps, it achieved TSS = 0.75 ± 0.05, HSS = 0.69 ± 0.04, POD = 0.76 ± 0.06 and FAR = 0.34 ± 0.10, yielding an overall F1 $\sim 0.70 \pm 0.04$. As compared to e.g. XGBoost, random forests' ensemble captures $\sim 11\%$ more true SEP events, lifting F1 (∼1%), TSS (∼11%) and HSS (∼1%), but at the cost of $\sim 7\%$ more false alarms. Variance is lower too (∼2%), hinting at slightly steadier behavior across cross-validation splits. Random forests show a clear improvement over the SWPC legacy probabilistic baseline. The model has been validated against the independent SEPVAL event sample, and the performance scores were POD = 0.75 and FAR = 0.22. Permutation importance scores reveal that flare soft X-ray fluence and CME speed mostly affect the model's predictions.

Access to Model Data and Forecasts: The flare and CME dataset (SEP-labeled) without imputed gaps, as well as the model training and evaluation pipeline developed during this study, are available at: https://github.com/SolarML/SEP-ML

Limitations, Caveats and Discussion: Despite its strong skill, the random forests show four operational weaknesses: i) they are prone to over-fitting, with training scores ~ 90 % but validation $\sim$ 10–15 % lower, revealing high variance; ii) they depend on having both flare and CME inputs, since dropping either source pushes all metrics below service thresholds; iii) their performance degrades when gaps are median-filled (F1 falls from 0.70 to 0.63) underscoring sensitivity to data quality; iv) any alert is bounded by the telemetry lag of CME speed and width measurements (e.g. ~ 6 hours for SOHO/LASCO), limiting real-time lead time for prediction.

A.5. *Empirical model for Solar Proton Event Real Time Alert (ESPERTA) Model*

Model Developers and Relevant Citation: Monica Laurenza, Edward W. Cliver, Alan G. Ling, Tommaso Alberti, Mirko Stumpo, Simone Benella; Laurenza et al. (2009, 2018, 2024), Alberti et al. (2017, 2019), Stumpo et al. (2021) and Benella et al. (2023).

Summary: The ESPERTA model introduced a logistic regression approach (Laurenza et al. 2009) for the prediction of solar proton events (defined as $\geq S1$ in the NOAA scale) following $\geq M2$ flares, and using three solar parameters: flare heliolongitude, soft X-ray fluence, and radio fluence —originally at 1 MHz from WIND/WAVES (Bougeret et al. 1995). The Laurenza et al. (2024) ESPERTA upgrades include: a) an ML approach with stratified cross-validation (Stumpo et al. 2021); b) a binary classification algorithm to forecast the occurrence of $\geq S1$ events and $\geq S2$ ones (defined as those reaching a peak flux of ≥ 100 pfu), to give an indication of the storm severity and c) replacement of space-based radio data with ground-based low-frequency observations (30 MHz from LOFAR) to allow for real-time operations. The upgraded ESPERTA maintains or improves performance compared to earlier versions, with POD up to 79% for $\geq S2$ events. Note here that NOAA categorizes solar radiation storms using the NOAA Space Weather Scale[13] on a scale from S1 - S5.

Model Description: Original ESPERTA uses logistic regression to estimate the probability of $\geq S1$ SPE occurrence based on the three input parameters. Predictions for $\geq S1$ events are issued 10 minutes after the $\geq M2$ soft X-ray flare peak time. The upgraded ML version of ESPERTA within a supervised learning framework, provides forecasting also for $\geq S2$ events at the time of $\geq S1$ threshold crossing. The $\geq S2$ proton events were identified for the period

[13] https://www.swpc.noaa.gov/noaa-space-weather-scales

Table 10: Model, Input and Output Specification Table for the ESPERTA model.

Model	
Type	Logistic Regression
Complexity	12
Input	
Shape	Point data (0D)
Type	Soft X-ray, Flare Location, Space-Based or Ground-Based Radio
History	23 years (1995–2017)*
Diversity	989 samples
Imbalance	0.1 positive
Sample Size	44 bytes
Sample Coverage	0 hours
Output	
Prediction	Classification, Probability
Type	Triggered
Forecast Window	7 hours
Comments: *Extended with recent data where available. **Average for S1, since the forecast window is ≈ 6–8 hours for ≥ $S1$ and 1.7–4 hours for ≥ $S2$.	

1995–April 2017, extending the list in Laurenza et al. (2018). A supervised ML approach was then applied, treating the ≥ $S1$ and ≥ $S2$ events as separate classes. The model was calibrated by determining the optimal threshold that maximizes the Critical Success Index (CSI), representing a balance between maximizing the POD and minimizing the FAR. The optimal thresholds were found to be 0.36 and 0.37 for ≥ $S1$ and ≥ $S2$ events, respectively.

Inputs: The three ESPERTA features are: a) flare heliolongitude; b) time-integrated soft X-ray flux (1–8$\mathring{A}$, GOES); c) time-integrated radio flux (originally 1 MHz WIND/WAVES, but can be replaced with 30 MHz LOFAR).

Outputs: The output of the model, based on a ≥ $M2$ flare trigger, are: probability of ≥ $S1$ SPE occurrence 10 minutes after flare peak and probability of ≥ $S2$ SPE occurrence after S1 onset. The probability is converted to a binary label (True/False) using the aforementioned optimal thresholds, informing us whether a flare will produce an SEP event.

Model Configuration: A logistic regression model with three features with a threshold optimized via CSI and for the upgraded ESPERTA (Laurenza et al. 2024) a supervised ML classifier for two-class prediction (≥ $S1$ and ≥ $S2$ severity levels).

Model Validation and Results: The model has been validates by using N-1 observations in the training set and 1 event in the test set and repeated this for N-1 times. The optimization led to the following scores. For the upgraded ESPERTA and for ≥ $S2$ events: theoretical POD = 0.88 (operational 0.79), FAR = 0.32 and median warning time ≈ 2 hours. Using 30 MHz ground-based data for ≥ $S1$ events (between 1995–2017) the upgraded ESPERTA achieved POD = 0.69 and FAR = 0.33.

Access to Model Data and Forecasts: Original ESPERTA uses WIND/WAVES and GOES data (publicly available from NASA/NOAA). Ground-based 30 MHz data available from LOFAR (van Haarlem et al. 2013) archives. The whole ESPERTA dataset, although not publicly available yet, can be obtaining by contacting the authors of Laurenza et al. (2024).

Limitations, Caveats and Discussion: Class imbalance —especially for ≥ $S2$ events— affects model calibration and FAR, and there is limited real-time availability of global low-frequency radio coverage as the single LOFAR station limits ESPERTA to 24-hour operation.

A.6. *University of MAlaga Solar particle Event Predictor (UMASEP) Model*

Model Developers and Relevant Citation: Marlon Nunez and Daniel Paul-Pena; Núñez (2011).

Summary: The UMASEP (University of Málaga Solar Energetic Proton) model introduces a dual-model ML framework for forecasting > 10 MeV SEP events by distinguishing between well-connected and poorly connected magnetic configurations between the Sun and Earth. The first model identifies well-connected events by detecting

Table 11: Model, Input and Output Specification Table for the UMASEP model. Please note that due to the authors of this work not participating in this effort, the values of the below table are estimates derived from Núñez (2011).

Model	
Type	Decision Tree, Linear Regression, Ensemble
Complexity	20
Input	
Shape	Time Series (1D)
Type	Soft X-Ray, Proton Flux
History	19 years (1987-2006)
Diversity	166 samples
Imbalance	0.45 positive
Sample Size	10,000 bytes
Sample Coverage	24 hours
Output	
Prediction	Classification, Regression
Type	Continuous
Forecast Window	5.17 hours

temporal correlations between GOES soft X-ray and differential proton flux time series, empirically estimating magnetic connectivity and associating flares of C7 or greater class, with potential proton enhancements. The second model targets poorly connected events using an ensemble of nonlinear regression trees trained on historical proton flux profiles to recognize patterns preceding gradual flux increases. An additional high-level analysis module filters inconsistent forecasts and estimates the expected intensity during the first 7 hours of the predicted event. Validated on SC 22–23, UMASEP achieved POD = 0.81, FAR = 0.40 and an average warning time of 5 hours and 10 minutes (1 hour and 5 minutes for well-connected and 8 hours and 28 minutes for poorly connected events), outperforming earlier automatic forecasters such as ESPERTA.

Model Description: The UMASEP model is based on an empirical dual-predictor design that analyzes the magnetic connectivity between the Sun and Earth to forecast $\geq$ 10 MeV SEP events. It includes two complementary components: one for well-connected events, which identifies time correlations between GOES soft X-ray flux and proton flux increases to infer flare–particle linkages, and another for poorly connected events, which applies a set of nonlinear regression trees to proton flux time series to detect gradual rises indicative of eastern limb or backside origins. Both predictors operate continuously and independently, generating forecasts when predefined thresholds of correlation or regression output are reached. A final decision module integrates both predictors, filters contradictory results, and estimates the expected proton intensity during the first 7 hours after onset. This architecture allows UMASEP to automatically detect the solar origin of each event type and provide early, interpretable warnings using only near-real-time GOES data.

Inputs: The inputs to UMASEP consist entirely of near–real-time measurements from GOES satellites, specifically the soft X-ray flux (0.1–0.8 nm channel) and differential proton flux in the $>$ 10 MeV energy range. These two time-series form the basis of both predictors within the model. For the well-connected predictor, UMASEP computes the evolving correlation between short time windows of X-ray and proton flux data to infer the magnetic linkage between the flare site and Earth. For the poorly connected predictor, the model uses only the proton flux data to train regression trees that recognize characteristic patterns preceding gradual proton enhancements. The temporal cadence of the inputs is typically 5 minutes, ensuring sufficient resolution for real-time SEP forecasting.

Outputs: The outputs of UMASEP are binary SEP occurrence forecasts indicating whether a $>$ 10 MeV proton event will occur, along with the estimated onset time and predicted intensity profile for the first 7 hours after detection. Each of the two predictors (well-connected and poorly connected) independently issues a probability-based SEP warning, which is then integrated by the decision module into a single operational forecast. The model also identifies the most likely solar origin—classifying events as well- or poorly connected based on the timing and strength of X-ray–proton correlations. UMASEP's forecasts are generated automatically and continuously in real-time, producing both the categorical SEP warning and quantitative estimates of expected proton flux intensity.

Model Configuration: The UMASEP model operates as a real-time forecasting system built around two independently configured predictors. The first predictor, for well-connected events, continuously computes correlation coefficients between short sliding windows of soft X-ray and proton flux to detect magnetically connected flare–particle pairs. The second predictor, for poorly connected events, uses an ensemble of nonlinear regression trees trained on historical GOES proton flux profiles from SCs 22 and 23. Both components are configured to update every 5 minutes, applying adaptive thresholds for correlation and flux variation that trigger a forecast when exceeded. A decision layer then combines the outputs, giving priority to well-connected predictions when both models issue simultaneous warnings. The entire framework was calibrated using a multi-cycle dataset and validated against NOAA's operational criteria for > 10 MeV SEP event detection.

Model Validation and Results: The UMASEP model was validated using data from SCs 22 and 23 (January 1986 – December 2009), encompassing 166 > 10 MeV SEP events listed in NOAA's catalog. Validation involved running the model in a simulated real-time mode using historical GOES X-ray and proton flux data. UMASEP achieved a POD of 0.81, a FAR* of 0.40, and an average warning time of 5 hours and 10 minutes. The model's separate predictors yielded mean lead times of 1 hour and 5 minutes for well-connected SEPs and 8 hours 28 minutes for poorly connected events. These results outperform previous automatic forecasters such as ESPERTA, particularly in detecting eastern-hemisphere events. The study also showed that UMASEP maintained stable performance across both SCs, demonstrating its robustness and operational readiness for real-time SEP forecasting.

Access to Model Data and Forecasts: All input data are publicly available through the NOAA/GOES archive and require no additional preprocessing beyond smoothing and normalization applied internally by the model. The model's performance can be found on the CCMC SEP Scoreboard by following this link: https://ccmc.gsfc.nasa.gov/scoreboards/sep/.

Limitations, Caveats and Discussion: A main limitation of UMASEP is its dependence on GOES satellite data availability and quality, which makes the model sensitive to telemetry gaps or delayed data streams that can interrupt real-time forecasting. The approach also assumes that soft X-ray–proton correlations accurately represent magnetic connectivity, which may not always hold for complex or multi-source events. While the dual-predictor architecture improves detection of both well- and poorly connected SEPs, it occasionally produces false alarms for proton flux fluctuations unrelated to solar flares. UMASEP's average warning time of about five hours is constrained by the onset of measurable X-ray or proton signatures, limiting its utility for extremely rapid SEP events. Additionally, the model does not incorporate data from coronagraphs, radio bursts, or CME kinematics, which could enhance prediction accuracy. Despite these caveats, UMASEP remains a reliable operational tool that balances interpretability and automation, demonstrating robust performance across two SCs and forming the foundation for later real-time forecasting systems such as UMASEP-10.

A.7. *University of Malaga predictor from Solar Data (UMASOD) Model*

Model Developers and Relevant Citation: Marlon Nunez and Daniel Paul-Pena; Núñez & Paul-Pena (2020).

Summary: The UMASOD (University of Málaga Solar Data Predictor) model introduced an interpretable ML framework to forecast > 10 MeV SEP events from solar flare and space-based radio burst data obtained from NOAA/SWPC event lists (updated every 30 min). Using a J48 decision tree (Panigrahi & Borah 2018), the model was trained on 502 events (75 SEP-producing and 427 non-SEP) spanning November 1997 to February 2014 ($\sim$ 16 years). Each event corresponds to a compact record ($\sim$ 0.001 MB) of flare and radio parameters such as soft X-ray flux, duration, and radio type III intensity and frequency range.

Model Description: UMASOD uses a decision tree (J48) classifier trained on flare and radio burst parameters from NOAA/SWPC event lists to forecast > 10 MeV SEP events. The model combines features such as soft X-ray peak and integrated flux, flare duration and rise time, heliographic coordinates, and radio burst intensity and frequency range to identify pre-SEP scenarios. During training, the algorithm found that the most relevant predictors were the integral of soft X-ray flux, flare rise time, and radio type III maximum frequency. The resulting tree provides interpretable decision paths for distinguishing SEP-producing from non-SEP events and was optimized using the CSI to balance POD and FAR*, achieving performance comparable to the ESPERTA model.

Inputs: The inputs to the UMASOD model consist of flare and radio burst parameters derived from NOAA/SWPC's Solar Edited Event Lists (updated every 30 minutes) and cross-referenced with the NOAA/NASA SEP list to label events as SEP or non-SEP. Each record includes flare information such as start, peak, and end times, heliolongitude and heliolatitude, X-ray peak flux (logarithmic scale), duration, and rise time, as well as radio burst characteristics

Table 12: Model, Input and Output Specification Table for the UMASOD model. Please note that due to the authors of this work not participating in this effort, the values of the below table are estimates derived from Núñez & Paul-Pena (2020).

Model	
Type	Decision Tree
Complexity	90
Input	
Shape	Time Series (1D)
Type	Soft X-Ray, Space-Based Radio
History	17 years (1997-2014)
Diversity	502 samples
Imbalance	0.14 positive
Sample Size	1,000 bytes
Sample Coverage	0.5 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	9.87 hours

like type (II–V), intensity, duration, frequency range, and integrated flux. The preprocessing step discretized flare coordinates into 12 regions, converted the X-ray flux scale to a linearized logarithmic form, and calculated additional variables such as the integral of soft X-ray flux, flare rise time, and product of soft X-ray and type III integrals, which served as combined predictive attributes. Events with a flare peak $\geq M2$ were retained, resulting in 502 labeled instances (75 SEP, 427 non-SEP) used to train the J48 decision tree.

Outputs: The output of UMASOD is a binary classification indicating whether a solar flare and radio burst event will result in a $>$ 10 MeV SEP occurrence. The model predicts either an SEP-producing (positive) or non-SEP (negative) outcome based on flare and radio parameters. The outputs are derived from decision tree rules optimized for operational use, providing interpretable "if–then" conditions that identify pre-SEP configurations using only solar data available at the time of flare and radio emission.

Model Configuration: The UMASOD model was implemented using the J48 decision tree algorithm within the Weka ML environment. The model was trained on 502 labeled events (75 SEP, 427 non-SEP) and optimized through 20-fold cross-validation, varying the minimum number of instances per leaf to maximize the CSI. The optimal configuration was found with eight instances per leaf, yielding a balance between Probability of Detection and False Alarm Ratio. Events with flare classes below M2 were filtered out, and the model trained exclusively on greater than or equal to M2 flare–associated cases. All computations were based on preprocessed flare and radio attributes, and the final decision tree, shown in the paper, reflects the interpretable rule-based structure generated by Weka for operational use.

Model Validation and Results: The UMASOD model was validated using historical flare and radio burst events spanning November 1997 to February 2014, employing 20-fold cross-validation during training and an independent evaluation of 104 flare-associated SEP events for performance assessment. Performance evaluation produced a POD of 0.70, a FAR* of 0.40, and an average warning time of 9 hours and 52 minutes, closely matching the empirical ESPERTA model. When optimized using the CSI, the model reached POD = 0.85 and FAR = 0.55 during cross-validation. These results confirm that UMASOD provides comparable accuracy to existing empirical predictors while maintaining interpretability and real-time operational capability based solely on solar flare and radio burst observations.

Limitations, Caveats and Discussion: A key limitation of the UMASOD model lies in its event-triggered and empirical design, which restricts forecasts to periods following flare and radio detections, preventing continuous monitoring of the solar environment. The class imbalance between SEP-producing (75) and non-SEP (427) events introduces bias toward negative predictions, though the use of the CSI helps mitigate this effect. While the decision-tree structure enhances interpretability, it may oversimplify nonlinear relationships among flare and radio parameters, limiting generalization to unseen or extreme events. The model also depends entirely on NOAA/SWPC event lists that

update every 30 minutes; any delay or data gap directly impacts its real-time applicability. Furthermore, the training data end in 2014, leaving performance under recent SCs unverified. Despite these caveats, UMASOD demonstrates that combining flare and radio burst parameters can yield operationally useful, physically interpretable forecasts for > 10 MeV SEP events, achieving skill scores comparable to more complex empirical systems such as ESPERTA.

A.8. *MS-SEP Model*

Model Developers and Relevant Citation: Mohammed AbuBakr Ali, Ali G. A. Abdelkawy, Abdelrazek M. K. Shaltout, and M. M. Beheary; Ali et al. (2025).

Table 13: Model, Input and Output Specification Table for the MS-SEP model.

Model	
Type	Random Forest
Complexity	52
Input	
Shape	Time Series (1D)
Type	Coronagraph, Soft X-Ray, Space-Based Radio
History	25 years (1997-2022)
Diversity	740 samples
Imbalance	0.108
Sample Size	416 bytes
Sample Coverage	6.5 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	4.6 hours
Comments: *These values are for the fixed frequency model. The values for the sweep frequency are a) Complexity: 29, b) Diversity: 534, c) Imbalance: 0.15, d) Sample Size: 232 and Forecast Window: 4.23 hours.	

Summary: This study developed an interpretable machine-learning framework to forecast > 10 MeV SEP events associated with M2.0 and stronger solar flares by integrating flare, space-based radio-burst, and CME data from NOAA/SWPC, NOAA/NASA and SOHO/LASCO catalogs. The system combines multi-source observations, including solar flare and radio bursts updated every 30 minutes and CME reports issued every 6 hours, covering 1997–2022. Random forest, decision tree, and SVM classifiers were trained and validated using both sweep-frequency and fixed-frequency radio datasets. The final dataset links each solar event to a compact feature record containing flare intensity, soft X-ray flux, CME speed and angular width, and type II/III radio-burst characteristics. Incorporating CME and radio parameters improved predictive skill, while restricting the analysis to stronger flares enhanced class balance. Nested cross-validation ensured robust and unbiased evaluation.

Model Description: This study implemented a binary classification framework to issue SEP and non-SEP warnings following $\geq M2.0$ solar flares with associated CME and radio-burst activity. The system integrates flare, CME, and radio-burst parameters into compact event-based records and evaluates multiple ML classifiers, including ramdom forest, Decision Trees, and linear and non-linear SVMs. Input features comprise flare soft X-ray flux and duration, CME speed and angular width, and type II/III radio-burst intensity and frequency characteristics to identify pre-SEP conditions. During training, CME kinematics and flare intensity consistently emerged as the most informative predictors. Model performance was assessed under imbalanced, balanced, and hybrid sampling strategies using nested cross-validation to ensure robust and unbiased estimates. Among the tested approaches, the random forest provided the most reliable and stable forecasting skill.

Inputs: The input dataset comprises solar flares associated with SEP events labeled as positive cases (80 SEP-producing flares) and flare-only events with concurrent CMEs and radio bursts labeled as negative cases (454 in the sweep-frequency set and 660 in the fixed-frequency set). Observations were compiled from NOAA/SWPC Solar

Event Lists for flare and radio-burst parameters, the SOHO/LASCO Coordinated Data Analysis Workshop (CDAW) catalog[14] for CME characteristics, and the NOAA/NASA ≥ 10 MeV proton flux list for SEP labeling, covering the period 1997–2022. Two complementary event-based datasets were constructed. The sweep-frequency dataset (534 events) includes features derived from Type II and Type III dynamic radio spectra together with flare and CME properties such as flare rise time, duration, soft X-ray flux and intensity, heliographic location, and CME speed and angular width. The fixed-frequency dataset (740 events) incorporates radio-burst intensity, duration, and flux measurements at multiple discrete frequencies, combined with the same flare and CME parameters. All variables were standardized using min-max scaling, and events with missing data were excluded to ensure consistency and reliable model training.

Outputs: The model produces a binary classification indicating whether a flare–CME–radio event is expected to generate a > 10 MeV SEP occurrence, labeling each case as SEP-producing (positive) or non-SEP (negative). Model performance is evaluated using standard skill metrics, including POD, FAR*, TSS, and HSS.

Model Configuration: The machine-learning framework was implemented in Python using the *sklearn* library and included random forest, decision trees, and linear and non-linear SVM classifiers. Model performance and hyperparameters were optimized using a nested cross-validation scheme with five outer and five inner folds to obtain robust and unbiased estimates. Hyperparameter tuning was conducted through randomized search, and class-weight adjustments were applied to mitigate dataset imbalance. All features were scaled using min-max normalization, and a fixed random seed ensured reproducibility. Model selection prioritized configurations that balanced detection capability and false alarms based on precision, Recall, POD, FAR, TSS, and HSS. Final performance statistics were computed as the mean and standard deviation across the outer folds.

Model Validation and Results: The framework was validated using historical flare, CME, and radio-burst events spanning 1997–2022, with performance assessed through nested cross-validation under imbalanced, balanced, and hybrid sampling conditions. Among the evaluated classifiers, the Random Forest consistently delivered the strongest performance across both datasets. For the sweep-frequency dataset, the model achieved a POD of 0.85 ± 0.08, a FAR of 0.30 ± 0.05, a TSS of 0.78 ± 0.07, a HSS of 0.71 ± 0.03, and an average warning time of approximately 5 hours. For the fixed-frequency dataset, corresponding values were POD $= 0.76 \pm 0.12$, FAR $= 0.31 \pm 0.08$, TSS $= 0.71 \pm 0.11$, HSS $= 0.67 \pm 0.06$, and an average warning time of about 4.5 hours. The results indicate stable generalization with no evidence of overfitting and improved detection capability compared with earlier empirical and machine-learning approaches. Feature importance analysis showed that CME speed and angular width were the dominant predictors, followed by flare intensity, soft X-ray flux, and key radio-burst characteristics.

Access to Model Data and Forecasts: All datasets and source files used in this study are publicly accessible through established solar data repositories. Solar flare and radio-burst data were obtained from the NOAA Space Weather Prediction Center (SWPC) Solar Event List, the CME data were retrieved from the SOHO/LASCO Coordinated Data Analysis Web (CDAW) catalog and the SEP event data were collected from the NOAA/NASA SEP list. The SolarML/SEP-ML codebase was adapted from Lavasa et al. (2021), applying modifications to the weighting scheme and restricting the analysis to four models: random forest, decision tree, SVM, and linear SVM. The adapted implementation is available at: https://github.com/SolarML/SEP-ML. Processed datasets and model scripts supporting this study are available upon reasonable request from the corresponding authors (Ali et al. 2025).

Limitations, Caveats and Discussion: The proposed framework focuses on SEP events associated with strong ($\geq M2.0$) flares, which excludes weaker-flare–driven events and restricts applicability to high-activity scenarios. The requirement for complete flare, CME, and radio-burst observations further reduces the available sample size and limits event diversity. Class imbalance between SEP and non-SEP cases remains an inherent challenge that may bias predictions despite mitigation strategies. Operationally, the model depends on external event catalogs, and delays in CME reporting—particularly the ~ 6 hour latency of SOHO/LASCO detections and manual CDAW updates—constrain real-time forecasting and shorten the effective warning window. These factors limit continuous monitoring capability and may affect generalization to unseen or extreme solar conditions.

A.9. *Classification and Regression Tree (CART) Model*

Model Developers and Relevant Citation: Soukaina Filali Boubrahimi, Berkay Aydin, Petrus Martens, and Rafal Angryk; Boubrahimi et al. (2017).

[14] https://cdaw.gsfc.nasa.gov/CME_list/

Table 14: Model, Input and Output Specification Table for the CART model.

Model	
Type	Decision Tree
Complexity	61
Input	
Shape	Time Series (1D)
Type	Soft X-ray, Energetic Protons
History	16 years (1997-2013)
Diversity	94 samples
Imbalance	0.5
Sample Size	59,451 bytes
Sample Coverage	10 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	0 hours*
Comments: *A forecast window of 0 hours indicates that the model performs event-level classification rather than temporal forecasting. These models rely on the GSEP catalog to label whether an SEP event occurs, without predicting its onset time or lead interval. Consequently, the output reflects the presence or absence of an SEP event conditioned on the input observations, not a forward-looking warning horizon.	

Summary: The paper introduces a method for predicting $\geq$ 100 MeV SEP events using GOES satellite data, focusing on time series from both X-ray and proton flux channels. It uses a Vector Autoregression (VAR) model to capture cross-channel correlations, including interactions among proton channels and between X-ray and proton data. Features extracted from these time series are used to train interpretable decision tree models on a balanced dataset of SEP and non-SEP events. The results show that certain correlations, especially involving proton channel P6 and the long X-ray channel, are strong indicators of upcoming SEP events. The proposed method achieves similar accuracy to existing systems like UMASEP while offering clear interpretability.

Model Description: The model predicts > 100 MeV SEP events using multivariate time series data from GOES satellite proton and X-ray channels. It applies a VAR model to capture linear dependencies among time series, focusing on how proton channel fluctuations relate to past values of both themselves and the X-ray channels (short and long wavelength). Each time series window (spanning up to 30 hours before an X-ray flare) is represented by a feature vector of VAR coefficients, expressing how proton responses are influenced by earlier activity. These features are used to train interpretable Classification and Regression Tree (CART) models. The decision trees use splitting criteria based on Gini impurity or information gain to identify feature thresholds that best separate SEP and non-SEP classes. The model is trained and evaluated on a balanced dataset of 47 SEP and 47 non-SEP events, using stratified 10-fold cross-validation. Results highlight that features such as the correlation between proton channel P6 and the long X-ray channel are among the most predictive for SEP event occurrence.

Inputs: The model takes as input multivariate time series data consisting of GOES X-ray and proton flux measurements. Specifically, it uses two X-ray channels—short (0.05–0.3 nm) and long (0.1–0.8 nm) —along with six proton channels: P6 and P7 from the EPS instrument (covering 80–500 MeV) and P8 to P11 from the High Energy Proton and Alpha Detector (HEPAD; Hanser 2011) instrument (covering 350 MeV to > 700 MeV). Each input sequence spans a fixed observation window, up to 30 hours before the onset of an X-ray event, sampled at a 5-minute cadence. These raw time series are then transformed into feature vectors using a VAR model that captures the dependencies of each proton channel on its own past values and those of the X-ray channels.

Outputs: The output of the model is a probability of SEP occurrence based on a flare trigger. The probability is converted to a binary label (True/False), informing us whether a flare will produce an SEP event.

Model Configuration: The model is configured as a CART decision tree classifier, trained using feature vectors derived from VAR applied to multivariate time series data. Two key parameters are tuned: the observation window span (ranging from 3 to 30 hours) and the VAR lag order (tested for values 1, 3, 5, 7, and 9), which controls how far back in time dependencies are modeled. The best performance was achieved with a 30-hour span and a lag of 5, balancing model complexity and predictive accuracy. The tree uses either Gini impurity or information gain as the splitting criterion, and training is done using stratified 10-fold cross-validation to ensure balanced class distribution and robust evaluation.

Model Validation and Results: The model is validated using stratified 10-fold cross-validation on a balanced dataset of 47 SEP and 47 non-SEP events, ensuring equal representation of both classes in each fold. Performance is assessed using standard classification metrics including accuracy, precision, recall, F1-score, and AUC. The best results are achieved with a 30-hour span and lag value of 5, using information gain as the splitting criterion. Under this configuration, the model reaches an Accuracy of 0.78, Precision of 0.86, POD of 0.73, F1-score of 0.82, and AUC of 0.77, indicating strong predictive power. These results are comparable to or slightly better than existing systems like UMASEP, with the added advantage of interpretability through decision tree rules.

Limitations, Caveats and Discussion: The model relies on historical correlations and may not generalize well to unseen solar conditions or rare event types. Missing data, especially in channels P6 and P7 during GOES-12, poses a risk of bias despite balancing. It also assumes a flare-based trigger, limiting applicability to flare-independent SEP events.

A.10. *Random Hivemind (RH) Model*

Model Developers and Relevant Citation: Patrick M. O'Keefe, Viacheslav Sadykov, Alexander Kosovichev, Irina N. Kitiashvili, Vincent Oria, Gelu M. Nita, Fraila Francis, Chun-Jie Chong, Paul Kosovich, Aatiya Ali, Russell D. Marroquin; O'Keefe et al. (2024).

Table 15: Model, Input and Output Specification Table for the RH model.

Model	
Type	Neural Networks, Ensemble
Complexity	202**
Input	
Shape	Point Data (0D)
Type	Soft X-ray, Flare Location
History	15 years (2002-2017)
Diversity	18,311 samples
Imbalance	0.0035 positive
Sample Size	48 bytes
Sample Coverage	Varies (length of the associated flare)
Output	
Prediction	Triggered
Type	Classification
Forecast Window	0 hours
Comments: *No window used in this study as it associates flares with SEP events. **The values in this table concern the RHv2 model.	

Summary: In this study, the considered problem is whether the particular soft X-ray flare event on the Sun is associated with the $\geq$ 10 MeV $\geq$ 10 pfu SEP event sometime in the future (with no particular forecasting time window). The ML model implemented is the Random Hivemind (RH) model, which represents the ensemble of individually-trained NNs, each considering a randomized set of features and voting proportionally to their importance. The input to the model is the soft X-ray properties of solar flares coming from the Temperature and Emission measure-Based Background Subtraction (TEBBS; Ryan et al. 2012, 2013) algorithm (Sadykov et al. 2019), along with their locations. The RH has been compared to the conventional NN approach (by keeping about the same architecture as

for the ensemble members) and to the committee approach (identical NNs trained individually for a majority vote). It was demonstrated that RH has a comparable or better performance with respect to the models it has been compared to, has a lesser spread of the scores for individual train-validation subsets, and captures almost all SEP events (making it a promising solution for all-clear predictions).

Model Description: The Random Hivemind (RH) model represents the ensemble NN model. Unlike the traditional approach, where the identical NN architectures are trained individually and issue a majority vote, the RH utilizes a) a randomized set of features propagating into each ensemble member, b) an adaptive learning rate that depends on the importance of features propagating to the ensemble member, and c) the output vote weighted for each ensemble member based on the importance of the features propagated into it. The basis network architecture includes two linear layers (10 neurons each), one dropout layer (20% rate), and one linear layer (2 neurons). The feature importances are estimated using the combination of χ^2 and mutual information gain statistics. Two versions of RH are considered using different numbers of features as inputs, and different approaches for progressing learning rates.

Inputs: The model utilizes the soft X-ray properties of solar flares computed from the 1-8 $\AA$ and 0.5-4 $\AA$ emission observed by the GOES XRS instrument. The properties are computed using the updated TEBBS algorithm. These properties include peak temperature, peak emission measure, background-subtracted flare class, and flare duration. The times of the temperature, emission measure, and soft X-ray flux peaks relative to the flare start and end times are also computed. Together with the flare coordinates, this results in 12 features per flare. For the flare-SEP association, the list of Solar Proton Events Affecting the Earth's Environment has been used and can be found here: https://www.ngdc.noaa.gov/stp/space-weather/interplanetary-data/solar-proton-events/SEP%20page%20code.html.

Outputs: The model produces a binary prediction of whether a particular flare is associated with the $\geq$ 10 MeV $\geq$ 10 pfu SEP event sometime in the future.

Model Configuration: Overall, one can configure the number of ensemble members, the number of features propagating to the estimators, the way of progressing the learning rate and other. One of the advantages of the model is that the newly trained ensemble member can be added without adjusting the previously trained members.

Model Validation and Results: The model has been validated using a variety of metrics, including widely-used TSS, HSS, and AUC. The model has been tested against the conventional NN of approximately the same architecture as RH ensemble members and having all features as inputs, and the committee ensemble of the identical networks. The performance of RH was found to be comparable or better than the competing approaches (TSS = 0.944 ± 0.023 and HSS = 0.168 ± 0.013 for RH version 2). The RH also typically demonstrated lower standard deviations for the scores, resulting in being less dependent on the particular train-test subdivision. In addition, the RH resulted in a very few false-negative predictions, demonstrating that it captures almost every SEP event, which would be desirable for all-clear purposes.

Access to Model Data and Forecasts: The SPE catalogs developed during this study are archived at the Solar Energetic Particle Prediction Portal which can be found here: https://sun.njit.edu/SEP3/datasets.html.

Limitations, Caveats and Discussion: This work has several important limitations, including a) a limited-span dataset that included only 64 unique SEP events; the validation on a larger dataset is desirable, and b) the non-operational nature of the currently implemented TEBBS algorithm, which requires the presence of the entire soft X-ray profile of the solar flare before producing the flare properties.

A.11. *Survival SEP (SSEP) Model*

Model Developers and Relevant Citation: India Jackson, Petrus Martens; Jackson & Martens (2024a) and Jackson & Martens (2024b).

Summary: The Survival SEP (SSEP) model applies survival analysis techniques to estimate the time-to-detection of SEP events following solar flares, using flare latitude, longitude, and GOES class as input features. Built on a curated dataset of flare-associated SEP events, the model implements Kaplan–Meier estimation and Cox PH modeling, with additional evaluation of survival trees and random survival forests. The output is a survival function $S(t)$, representing the probability that an SEP event has not occurred by time t.

Model Description: Five feature sets were tested, including combinations of flare latitude, longitude, and GOES class, as well as subsets selected based on Cox PH significance. No NN was used; instead, classical survival analysis models and tree-based ensembles were evaluated. Grid search and 5-fold cross-validation were used for hyperparameter tuning, although the models themselves contain no trainable weights.

Table 16: Model, Input and Output Specification Table for the SSEP model.

Model	
Type	Forest, Decision Tree
Complexity	300
Input	
Shape	Point Data (0D)
Type	Flare Location
History	31 years (1986-2017)
Diversity	293 samples
Imbalance	0.9044
Sample Size	47 bytes
Sample Coverage	0 hours
Output	
Prediction	Probability
Type	Triggered
Forecast Window	22.46 hours*
Comments: *The forecast window is variable, defined by the model's prediction of the time until SEPs exceed 10 MeV following a solar flare. The average forecast window is 22.46 hours.	

Inputs: The model uses point-based input data, where each solar flare event is independently represented as a set of features (time-to-detection, longitude, latitude, and GOES class) with no temporal or spatial relationship between samples. This structure qualifies as Point Data (0D features). The physical quantity represented is energetic protons detected at ≥ 10 MeV. The dataset spans 31 years, from 1986 to 2017, and includes a total of 293 labeled flare events: 265 positive cases (flare followed by an SEP) and 28 negative cases (flare not followed by an SEP), yielding a class imbalance of approximately 9.56% negative cases (or 90.44% positive). Each event consists of 5 features, resulting in a data size of approximately 40 bytes per sample. The time coverage of a single input sample, defined as the duration between flare onset and SEP detection, ranges from 89 minutes to 5,886 minutes across the dataset.

Outputs: Each trained survival model outputs a survival function $S(t)$, which indicates the probability that an SEP event has not occurred at a given time post-flare. The random survival forest version enhances interpretability and captures non-linear relationships in the flare–SEP timing data.

Model Configuration: The models were implemented in Python using *scikit-survival* and *sklearn*. Hyperparameters were tuned using *GridSearchCV*, but no learnable weights exist in these models. For the random survival forests, 300 estimators were used. Tree depth and split criteria were optimized for performance via log-rank test statistics.

Model Validation and Results: Performance was assessed using the concordance index (C-index) on held-out validation sets. The best-performing Cox PH model achieved a C-index of ~ 0.82, indicating strong predictive capability for time-to-detection of SEPs following solar flares. Feature importance analysis identified flare longitude as the most influential predictor, followed by GOES class and flare latitude. Models were validated using 5-fold cross-validation to ensure generalizability, and survival curves were compared between SEP and non-SEP events to evaluate separation. Survival tree and random survival forests methods yielded similar trends in predictor importance, with added benefits in capturing nonlinear interactions. Results support the viability of survival analysis as a practical forecasting framework for operational space weather applications.

Access to Model Data and Forecasts: All data and code are publicly available. The model dataset is hosted on Harvard Dataverse at https://doi.org/10.7910/DVN/GXY9MZ, and the codebase is maintained on GitHub at https://github.com/indiajacksonphd.

Limitations, Caveats and Discussion: This model assumes that SEP onset is directly related to flare timing, which excludes CME-only or shock-driven SEP events. Additionally, the flare-to-SEP association window may introduce uncertainty due to event overlap. Future work may expand to include CME parameters and solar wind context for increased accuracy.

A.12. *SEP-C Model*

Model Developers and Relevant Citation: Jesse Torres, Philip K. Chan, Lulu Zhao, and Ming Zhang; Torres et al. (2022).

Table 17: Model, Input and Output Specification Table for the SEP-C model.

Model	
Type	Neural network
Complexity	780
Input	
Shape	Point Data (0D Features)
Type	Proton Flux, Coronagraphs*, Solar Wind
History	21 years (1996-2017)
Diversity	20,210 samples
Imbalance	0.0046 positive
Sample Size	100 bytes
Sample Coverage	2 hours
Output	
Prediction	Classification, Probability, Regression
Type	Continuous
Forecast Window	24 hours
Comments: *This study uses properties of CMEs (Point Data) derived by the SOHO/LASCO coronagraphs.	

Summary: The SEP-C model is a NN based on characteristics of CMEs and Type II radio waves from the CDAW catalog, solar wind speed, and sunspot number, to forecast whether CMEs would lead to SEP events.

Model Description: The NN is a Multi-Layer Perceptron (MLP) with one hidden layer of 30 units and Rectified Linear Unit (ReLU) activations. The loss function is binary cross entropy with L2 regularization to reduce overfitting.

Inputs: The input features are based on characteristics of CMEs, Type II radio waves, and sunspot number. The basic CME features include linear speed, width, acceleration, second order speed initial, second order speed final, second order speed at 20 solar radii, central position angle, measurement position angle. Extended features include the number of CMEs with the past month and past 9 hours, the maximum speed of all CMEs within the past day, the number of CMEs with a speed greater than 1,000 km/s, $v * log(v)$ (where v is the linear speed), halo, and particle intensity based on Diffusive Shock Acceleration (Drury 1983). Other features are the solar wind speed at Earth, area in the spectrogram (duration $\times$ frequency range) of a Type II burst, sunspot number, and empirical SEP intensity prediction formula based on Richardson et al. Richardson et al. (2018). The imbalance ratio is about 1:300 and by varying oversampling a 1:3 ratio was found to be desirable.

Outputs: The output of the model is a score, which can roughly be interpreted as a probability, of whether a CME is associated with an SEP event. The natural logarithm of proton intensity 30 (or 60) minutes in the future.

Model Configuration: Training the model uses a learning rate of 0.1, momentum coefficient of 0.9, batch size of 200, and an L2 regularization term of 0.1. The model is allowed to train for up to 2,000 iterations before stopping unless the loss does not decrease by 10^{-4} within 10 consecutive iterations.

Model Validation and Results: This study indicates that the model with all the features can achieve 0.906 in TSS, 0.245 in HSS, and 0.246 in F1. The features were divided into 5 groups (speed, size, location, history, and others) and found that speed-related and other (e.g. Type II burst, sunspot) features are relatively more important. For forecasting intensity, our study indicates that this model can achieve 0.379 in Mean Absolute Error (MAE) for 30-minute forecast and 0.599 in MAE for 60-min forecast. For forecasting the start of SEP events exceeding 10 pfu intensity threshold of $\geq$ 10 MeV protons, periods of advanced and extended warnings are incorporated. The SEP-C model can achieve 0.76 in F1 for 30-minute forecast and 0.85 in F1 for 60-minute forecast.

Access to Model Data and Forecasts The CME list used in this study can be found in https://cdaw.gsfc.nasa.gov/CME_list/ and https://cdaw.gsfc.nasa.gov/CME_list/radio/waves_type2.html and the repository with the data and code can be found here https://doi.org/10.5281/zenodo.12832882.

Limitations, Caveats and Discussion: Most false alarms occur when CMEs are not fast and large. Future work will continue to reduce the false alarms by adding more features, particularly those that can distinguish SEPs with features that are similar to non-SEPs, such as low speed. The CME speed and size data from the CDAW catalog contain observation limitations such as projection effects. Including properties of CME size, location, and speed listed in the DONKI database can reduce the false alarm rate. The data in this study cover a partial SC. Particularly, the test data set in the evaluation is near the solar activity maximum. Extending the data to cover two SCs could help improve the model.

A.13. *Custom Architecture Neural Network (CANN) Model*

Model Developers and Relevant Citation; Viacheslav M. Sadykov, Alexander G. Kosovichev, Irina N. Kitiashvili, Vincent Oria, Gelu M. Nita, Egor Illarionov, Patrick M. O'Keefe, Yucheng Jiang, Sheldon H. Fereira, Aatiya Ali; Sadykov et al. (2021).

Table 18: Model, Input and Output Specification Table for the CANN model.

Model	
Type	Neural Network
Complexity	1,243
Input	
Shape	Point Data (0D)
Type	Soft X-ray, Proton Flux, Magnetic Fields and Ground-Based Radio
History	9 years (2010–2019)
Diversity	2,288 samples
Imbalance	0.0294
Sample Size	1,244 bytes
Sample Coverage	24 hours
Output	
Prediction	Classification, Probability
Type	Continuous
Forecast Window	24 hours

Summary: The daily whole-Sun binary prediction of SPEs ($\geq$ 10 MeV, $\geq$ 10 pfu) is considered, along with the related all-clear problem. The network is a custom-architected NN that utilizes the magnetic field properties of the solar ARs, the statistical properties of the preceding soft X-ray and proton fluxes, as well as records of solar radio bursts, as an input. The model has been evaluated on the SC 24 data, and demonstrated the performance comparable to or better than the daily probabilistic SWPC NOAA forecasts, especially in situations when missing the events is undesirable (all-clear regime).

Model Description: The ML model considered in this work is a custom-architected NN. The key idea of the custom architecture is to pre-process the magnetic field properties of individual ARs in the identical way first (in so-called AR blocks), then sum up the output from these blocks, and propagate it into the fully-connected part to concatenate with the whole-Sun statistical features from soft X-ray and proton flux observations and daily counts of radio bursts. Since each AR block shares the same weights and biases updated synchronously during the network training, the network has significantly fewer free parameters compared to the fully-connected analog, which makes it less prone to overfitting. The network implemented by Sadykov et al. (2021) has 30–15–8–4–2 neurons for AR blocks and 21–15–10–5–2 for the main network, resulting in 1243 free parameters.

Inputs: The model takes the median SHARP (Bobra et al. 2014) properties of 10 solar ARs with the largest unsigned magnetic flux (including AR locations and quality parameters) as inputs into the AR blocks. The ARs that are rotated behind the 68-degree longitude are assumed to have their SHARP properties unchanged for 11 days, and are propagated with the Carrington rotation rate. The model then concatenates the AR block outputs with daily soft

X-ray properties (mean, standard deviation, median, minimum, maximum), daily $\geq$ 10 MeV proton flux properties (mean, standard deviation, median, minimum, maximum, last value), and daily counts of type II, III, and IV radio bursts.

Outputs: The model produces a daily binary and probabilistic forecast for $\geq$ 10 MeV $\geq$ 10 pfu proton events.

Model Configuration: While a certain model architecture has been considered by Sadykov et al. (2021), the network can be adjusted with respect to any additional AR or whole-Sun properties (requires the model retraining).

Model Validation and Results: The model has been trained and tested on the SC 24 proton events. The train data set contained 2,222 non-SPE days and 66 SPE days (the days when the flux of $\geq$ 10 MeV protons has exceeded 10 pfus). These numbers are 1178 and 35 for the test data set, respectively. The model has been evaluated using the standard metrics of TSS, two variations of the HSS, AUC, and the new metric of weighted TSS introduced in Sadykov et al. (2021). The results have been compared with the performance of the SWPC NOAA daily probabilistic models. It was found that, overall, ML-driven prediction outperforms SWPC NOAA forecasts in the all-clear regimes (when missing an SEP is undesirable), and has a competitive TSS = 0.82 ± 0.01 and HSS = 0.38 ± 0.03 overall.

Access to Model Data and Forecasts: The SPE catalogs and the Jupyter notebooks developed during this study are archived at the Solar Energetic Particle Prediction Portal, which can be found here: https://sun.njit.edu/SEP3/datasets.html.

Limitations, Caveats and Discussion: The major limitation of the current work is the training and test data set sizes, which have included only the SC 24 data (and, therefore, a very limited number of unique SEPs). Other limitations include the consideration of relatively weak S1 NOAA events only ($\geq$ 10 MeV $\geq$ 10 pfu), reliance on the science-quality SHARP and GOES data instead of the operational data streams.

A.14. *SEP-E Model*

Model Developers and Relevant Citation: Jesse Torres, Philip K. Chan, Lulu Zhao, and Ming Zhang; Torres et al. (2025).

Table 19: Model, Input and Output Specification Table for the SEP-E model.

Model	
Type	Neural network
Complexity	1,530
Input	
Shape	Time Series (1D)
Type	Proton Flux, Electron Flux
History	7 years (1995-2002)
Diversity	517,769 samples
Imbalance	0.0103
Sample Size	200 bytes
Sample Coverage	2 hours
Output	
Prediction	Classification, Probability, Regression
Type	Continuous
Forecast Window	0.75 hours
Comments: *This value reflects the average of the forecast window range of 30 - 60 minutes.	

Summary: The SEP-E model is a NN based on electron and proton intensities to forecast proton intensities 30 (or 60) minutes in the future every 5 minutes.

Model Description: The model is an RNN with one hidden layer of 30 gated recurrent units. Mean squared error (MSE) is used for the loss function.

Inputs: Two hours of electron intensities from the $\geq$ 0.25 and $\geq$ 0.67 MeV channels and proton intensities $\geq$ 10 Mev channels at 5-minute intervals. Input to the model is the time series of natural logarithm of measured particle

intensities in pfu. solar X-ray emission data have also been used in the SEP-E model, but they did not help improve the prediction performance.

Outputs: The natural logarithm of proton intensity 30 (or 60) minutes in the future.

Model Configuration: Weights are updated using the Adam optimizer, and up to 1,000 iterations are allowed unless the network converges before then. The NN converges if the loss function does not change by more than 10^{-4} over 20 iterations.

Model Validation and Results: For forecasting intensity, our study indicates that our model can achieve 0.379 in MAE for 30-minute forecast and 0.599 in MAE for 60-minute forecast. For forecasting the start of SEP events exceeding 10 pfu intensity threshold of ≥ 10 MeV protons, periods of advanced and extended warnings are incorporated. SEP-E can achieve 0.76 in F1 for 30-minute forecast and 0.85 in F1 for 60-minute forecast.

Access to model data and Forecasts: The code and data for this work can be found here: https://doi.org/10.5281/zenodo.12832882.

Limitations, Caveats and Discussion: The data in this study cover a partial SC. Particularly, the test data set in the evaluation is near the solar activity maximum. Extending the data to cover two SCs could help improve the model.

A.15. *Space Radiation Intelligence System (SPRINTS) Model*

Model Developers and Relevant Citation: Alec Engell, Brianna Maze, Harold Farmer; Engell et al. (2017).

Table 20: Model, Input and Output Specification Table for the SPRINTS model.

Model	
Type	Neural Network
Complexity	5,401*
Input	
Shape	Point Data (0D)
Type	Soft X-ray, Flare Location
History	31 years (1986–2017)
Diversity	2,263 samples
Imbalance	0.067 *
Sample Size	32 bytes
Sample Coverage	0 hours
Output	
Prediction	Classification, Probability
Type	Triggered
Forecast Window	24 hours **
Comments: *These values correspond to the 10_10_24 model **SPRINT can have prediction windows that range from 0 to 96 hours, the results presented in this document is for the 24 hours prediction model.	

Summary: The Space Radiation Intelligence System (SPRINTS; Figure 5) is a modeling framework for data-driven forecasting of solar-driven events. It applies event catalogs and databased observations including flares, SEPs, CMEs, and radio bursts as well as associated catalogs such as flare events that are associated to SEP events. The developed SPRINTS SEP forecasting model uses an MLP model based on flare parameters including flare flux, flare fluence, flare decay phase, flare long/short X-ray ratio, and flare longitude.

Model Description: The MLP model is used for binary classification to distinguish between SEP and non-SEP events. Specifically, the model outputs a probability whether an SEP will occur within a given time window at a given proton energy channel. A separate MLP model is trained to predict SEP occurrences at user-defined proton energy channels (e.g., 1, 5, 10, 30, 50, 100 MeV), flux thresholds, and time resolutions (e.g., 12-hour) up to 96 hours in advance. One model is trained per combination of energy channel and time resolution, resulting in 20 independent MLP models.

Inputs: There are 252 flare-SEP events and 19,959 flare-only events that are cataloged in the SPRINTS database. To reduce the number of flare-only events used during training, flares that had a fluence below the minimum fluence observed in any SEP event were excluded. This filtering step reduced the flare-only set to 10,084 events, ensuring that the model was trained on more challenging negative samples. To further address the class imbalance between SEP and flare-only events, a random subset comprising 20% of the remaining flare-only events was selected for inclusion in the training set. This resulted in 2,017 flare-only events. Because the models were designed to predict whether an SEP would occur at a given proton energy channel and flux threshold, it was necessary to calculate whether each of the 252 events exceeded the threshold at each given energy channel. An SEP event was labeled as positive (1 label) for a given energy-threshold bin if it exceeded that threshold; otherwise, it was labeled negative (0 label). Since not all SEP events exceeded every threshold, the number of positive samples varied across bins, further reducing the positive samples down from the original 252 for each bin.

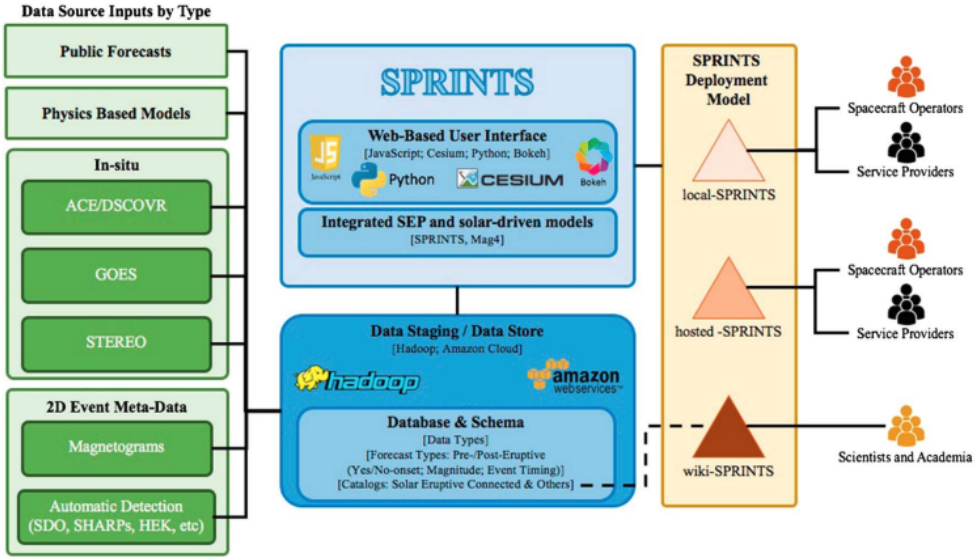


Figure 5: SPRINTS logical architecture organized by data sources, data staging/store, user interface, forecast models (integrated MAG4 and SPRINTS models), deployment models, and users.

Each model was trained using flare characteristics including X-ray flux, X-ray fluence, X-ray peak ratio (long vs. short channel), and flare longitude. Challenge events defined by CCMC were completely removed from the dataset and used for a hold-out test set. The remaining dataset was split into training and testing sets using a 90/10 ratio, with stratification to preserve the distribution of positive and negative samples. If the resulting test set did not include any SEP events, one SEP event was randomly moved from the training set to the test set to ensure representation. Feature normalization was performed using the *StandardScaler*[15] from Python's *sklearn* library, which standardizes features by removing the mean and scaling to unit variance. It's important to note the mean and standard deviation were computed from the training data alone and then applied to both training and testing sets to prevent data leakage.

Outputs: The model outputs a probability representing the likelihood that a given solar flare will produce an SEP event. This probability is converted into a binary classification (True/False) using a threshold of 0.5, indicating whether an SEP event is expected to occur given the flare metadata. The system is configurable to any desired GOES energy channel and flux threshold and is currently deployed in real-time for the following thresholds: 10 MeV at 10 pfu, 10 MeV at 40 pfu, 30 MeV at 10 pfu, 50 MeV at 10 pfu, 100 MeV at 1 pfu.

Model Configuration: A separate MLP model was trained for each energy channel and threshold requirement, resulting in a total of 20 models. To optimize performance, each model underwent a two-stage hyperparameter tuning process. Initially, randomized grid search with k-fold cross-validation using 3-folds was employed to efficiently explore a broad hyperparameter space. Based on the results of this preliminary search, a refined parameter space was defined

[15] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

for a full grid search, again using 3-fold cross-validation. The use of k-fold cross-validation helped ensure that the models were not overfitting to the training data. Hyperparameters tuned during this process included the number and size of hidden layers, activation functions, learning rate, and maximum number of training iterations. Because tuning was performed individually for each model, the final configurations varied across energy channels and time resolutions. Once the optimized parameters were found using cross-validation, the models were re-trained on the entire training dataset and saved.

Model Validation and Results: Model performance was evaluated using the HSS, POD, and False Detection Rate. HSS results of the models after training on the train set and evaluating on the hold-out test set range from 0.17 – 0.65 for the 24-hour time window across the different energy channels. It is important to note that as the forecast time window increases, in particular at 72 and 96 hours, the skill scores drop dramatically or exhibit large variability. This is primarily due to the substantial class imbalance between SEP and flare-only events at longer lead times, which results in poor representation of positive (SEP) samples for both training and evaluation.

Access to Model Data and Forecasts: Model outputs are publicly available on the CCMC SEP scoreboard https://ccmc.gsfc.nasa.gov/scoreboards/sep/ via the SPRINTS REST API.

Limitations, Caveats and Discussion: Because each MLP model is trained independently for a specific combination of temporal bin and proton energy/flux threshold, the resulting probability outputs are not inherently consistent across models. This means that the predicted probability of SEP occurrence at a higher energy channel may exceed that of a lower energy channel, even though such behavior may appear counterintuitive from a physical standpoint. This is a direct consequence of training a separate MLP model for each unique combination of temporal bin and energy/flux threshold. The disproportionate number of flare-only events relative to SEP events presented challenges during model training. Without proper handling, this imbalance can lead to biased models that underperform in predicting the minority class (SEPs). Careful attention was paid to create representative and balanced training and testing datasets. This includes strategies such as selective sampling of flare-only events and stratified data splitting, ensuring that each model is trained and evaluated on data that reflect the operational scenario as closely as possible.

A.16. *Time-Series Forest (TSF) Model*

Model Developers and Relevant Citation: Pouya Hosseinzadeh, Soukaina Filali Boubrahimi, and Shah Muhammad Hamdi; Hosseinzadeh et al. (2024a).

Table 21: Model, Input and Output Specification Table for the TSF model.

Model	
Type	Forest, Ensemble
Complexity	15,000
Input	
Shape	Time Series (1D)
Type	Energetic Protons
History	25 years (1986-2011)
Diversity	141 samples
Imbalance	0.0746 positive
Sample Size	11.52 bytes
Sample Coverage	5 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	0 hours*
Comments: *A forecast window of 0 hours indicates that the model performs event-level classification rather than temporal forecasting, similar to Table 14.	

Summary: This work introduces a time series data augmentation approach to improve SEP event prediction for $\sim$ 30, $\sim$ 60, and $\sim$ 100 MeV energy bands using GOES proton flux. Three time series classification models —Time

Series Forest (TSF; Figure 6b), ROCKET, and SHAPELET— are evaluated. The model performance is significantly enhanced by applying SMOTE, ADASYN, and Gaussian noise. Among all models, TSF consistently outperforms others, reaching up to 90% accuracy in the ∼ 100 MeV prediction task.

Model Description: The model uses time series classification techniques to distinguish between SEP and non-SEP events using GOES proton flux data. TSF, a random forest-based ensemble method (Figure 6c), is applied to 1D time series segments extracted from 5-hour observation windows. TSF randomly selects intervals from each time series and calculates statistical features such as mean, standard deviation, and slope. The model is trained using both real and synthetically generated SEP samples to balance the class distribution. This approach supports both binary classification (SEP vs. non-SEP) and hierarchical multi-class classification across energy levels (∼ 30, ∼ 60, ∼ 100 MeV).

Inputs: We used GOES proton flux time series data from three channels: P4 (∼ 30 MeV), P5 (∼ 60 MeV), and P6 (∼ 100 MeV), with 5-hour observation windows prior to solar flares. SEP event lists were extracted from the GSEP catalog, while non-SEP events were collected using the Heliophysics Event Knowledgebase (HEK[16]). The SEP labels correspond to the increase in energetic protons detected after solar flare activity. Multivariate and univariate versions of these time series were created and used as input for classification models, with data augmentation techniques applied to SEP samples to address class imbalance.

Outputs: The output of the model is a probability of SEP occurrence based on a flare trigger. The probability is converted to a binary label (True/False), informing us whether a flare will produce an SEP event.

Model Configuration: The best-performing model is the TSF, which uses an ensemble of decision trees built on random intervals of the time series. Each tree is trained on statistical features (mean, standard deviation, slope) extracted from selected time intervals of the GOES proton flux. Data augmentation is applied using Gaussian noise, SMOTE, and ADASYN methods to balance the SEP vs. non-SEP classes. Models are evaluated using k-fold cross-validation, and both univariate and multivariate time series inputs are considered. Hierarchical modeling predicts high-energy SEPs first (∼ 100 MeV), then medium (∼ 60 MeV), and finally low-energy (∼ 30 MeV) events.

Model Validation and Results: The TSF model achieved significant improvements with data augmentation. For ∼ 100 MeV SEP classification, accuracy increased from 70% to 90% using Gaussian noise. For ∼ 60 and ∼ 30 MeV, Gaussian yielded the best accuracy (90%). The model was evaluated using k-fold cross-validation with 10 folds for 100 MeV, and 7 folds for 60 and 30 MeV due to limited data. TSS and HSS scores increased to 0.8–0.9 with augmentation. SMOTE and ADASYN also improved performance, especially for 100 MeV classification.

Access to Model Data and Forecasts: All input data used in the TSF model are derived from publicly accessible heliophysics repositories. Time-series proton flux measurements are obtained from the GOES Space Environment Monitor (SEM) archives provided by NOAA National Centers for Environmental Information (NCEI), with SEP-producing events sourced from the GSEP catalog and corresponding non-SEP flare events obtained from the HEK. These publicly available time-series datasets enable reproducible SEP classification experiments for high-energy (∼ 100 MeV) event prediction under fixed 6-hour observation window configurations.

Limitations, Caveats and Discussion: A key limitation of SEP prediction is the extreme class imbalance due to the rarity of high-energy events, especially ∼ 100 MeV. Although data augmentation significantly improves classification performance, synthetic samples may not capture all physical characteristics of real SEP events. Additionally, performance varies by energy band (∼ 100 MeV classification remains more challenging than ∼ 30 or ∼ 60 MeV). While TSF achieves high accuracy, further improvement may require incorporating additional data modalities or physics-based constraints to better generalize to unseen SEP scenarios.

A.17. *Univariate Deep Merge (UDM) Model*

Model Developers and Relevant Citation: Pouya Hosseinzadeh, Soukaina Filali Boubrahimi, and Shah Muhammad Hamdi; Hosseinzadeh et al. (2024b).

Summary: This study presents a multimodal time series data fusion framework for predicting high-energy (∼ 100 MeV) SEP events by combining GOES proton flux data and solar EUV images. Six ML models are evaluated: two unimodal models—UTS (time series only) and Image (image only), and four fusion models (UFC, UDC, UDM, and USC). Among these, the Univariate Deep Merge (UDM) achieves the highest performance, reaching 0.80 Accuracy and 0.81 Precision under the balanced setting. The results highlight the importance of both temporal and spatial infor-

[16] https://www.lmsal.com/hek/

Table 22: Model, Input and Output Specification Table for the UDM model.

Model	
Type	Time Series Forest
Complexity	15,548
Input	
Shape	Time Series (1D), Vectors (1D)
Type	EUV Imagery, Energetic Protons
History	15 years (1997-2012)
Diversity	59 samples
Imbalance	0.2655 positive
Sample Size	11.52 bytes
Sample Coverage	6 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	0 hours*
Comments: *A forecast window of 0 hours indicates that the model performs event-level classification rather than temporal forecasting, similar to Table 14.	

mation for SEP classification, and show how optimal observation window sizes and image vector lengths significantly impact prediction accuracy.

Model Description: The best-performing model in this study is the UDM model, a multimodal data fusion architecture that integrates time series proton flux data and solar EUV image (converted to vectors). The model applies a deep learning-based strategy in which separate neural branches process each modality independently—one branch for 5-hour time series segments from the GOES P6 proton channel ($\sim$ 100 MeV), and another for 200-dimensional vectors extracted from EUV images using an autoencoder. These representations are merged using element-wise operations within the network to capture both shared and complementary features. The fused output is then passed through dense layers for final classification. UDM effectively captures temporal dynamics and spatial patterns, enabling robust prediction of high-energy SEP events.

Inputs: The model uses two data modalities: a) time series proton flux data from the GOES P6 channel ($\sim$ 100 MeV) over a 6-hour observation window preceding solar flares, and b) solar single images from SOHO's Extreme-ultraviolet Imaging Telescope (EIT; Delaboudiniere et al. 1995) 304Å channel, captured within 4 hours before the flare start time. The time series data captures temporal dynamics of energetic particle activity, while the EUV images are transformed into 200-dimensional latent vectors using an autoencoder to extract spatial features. SEP events are sourced from the GSEP catalog, and non-SEP events are selected from the HEK flare records with peak intensity $\geq C1.3$ that did not lead to SEP events.

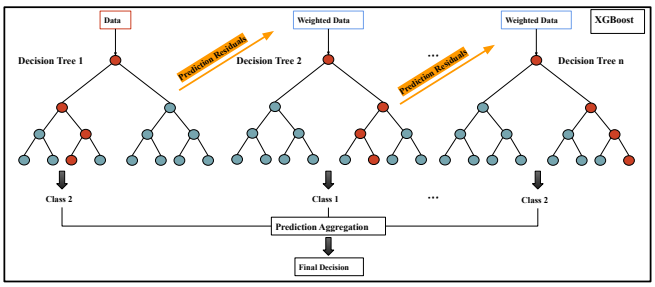
Outputs: The output of the model is a probability of SEP occurrence based on a flare trigger. The probability is converted to a binary label (True/False), informing us whether a flare will produce an SEP event.

Model Configuration: The UDM model consists of two parallel neural branches: one processes 6-hour GOES P6 time series using a sequence of dense and dropout layers, and the other processes 200-dimensional EUV image vectors extracted via an autoencoder. Features from both branches are merged using element-wise operations (e.g., addition), followed by fully connected layers and a softmax output for classification. The model is implemented using *TensorFlow*[17] and trained using the Adam optimizer. Hyperparameter tuning determined the optimal configuration, including 150 estimators for the base TSF model and 5-fold stratified cross-validation for both balanced and imbalanced data settings.

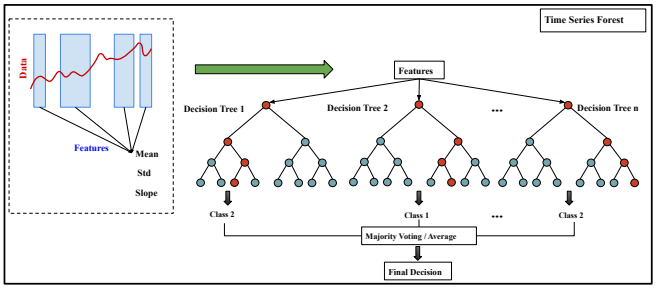
Model Validation and Results: Model performance was evaluated using 5-fold cross-validation on both balanced and imbalanced datasets. The univariate UDM model consistently outperformed all other models across multiple

[17] https://www.tensorflow.org/

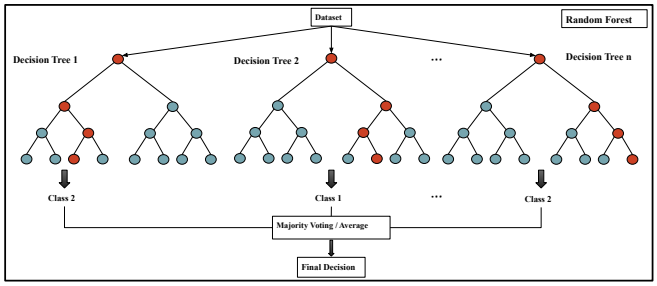
metrics. On the balanced dataset, UDM achieved 0.80 Accuracy, 0.79 F1-score, and 0.81 Precision. It also showed strong Recall, TSS, and HSS scores, indicating its reliability in detecting high-energy SEP events. The model's robustness was further confirmed through noise sensitivity analysis, where UDM maintained the highest F1-scores under varying levels of Gaussian noise. Compared to state-of-the-art baselines such as XGBoost, SVM, RF, and decision trees, UDM showed superior average F1 and TSS scores, validating its effectiveness for $\sim$ 100 MeV SEP prediction using multimodal fusion.



(a) XGBoost-style boosted decision-tree ensemble.



(b) Time Series Forest (TSF) pipeline.



(c) Random forest ensemble classifier.

Figure 6: Tree-based ensemble architectures used for SEP classification, including gradient boosting, Time Series Forest, and Random Forest models.

Access to Model Data and Forecasts:
The implementation of the UDM model, including preprocessing and training configurations, is available at https://github.com/pouyahosseinzadeh/High-Impact-SEP-Prediction---Space-Weather. All input data used in this study are derived from publicly accessible heliophysics repositories. Time-series proton flux measurements are obtained from the GOES SEM archives provided by NOAA NCEI, while solar EUV imagery is accessed via the SOHO instrument through the *Helioviewer*[18] platform. SEP event lists are sourced from the GSEP catalog, and non-SEP events are collected from the HEK. These multimodal inputs enable reproducible SEP classification experiments under fixed observation windows for high-energy ($\sim$ 100 MeV) event prediction.

Limitations, Caveats and Discussion: A key limitation of this study is the limited number of high-energy ($\sim$ 100 MeV) SEP events available, which restricts the size and diversity of the dataset. Although balanced and imbalanced settings were both evaluated, real-world scenarios involve extreme class imbalance that may impact generalization. The SOHO EUV images, used as part of the multimodal input, are infrequent and sometimes captured hours before flare onset, introducing temporal variability and potential misalignment with the time series data. Additionally, the model focuses solely on classification (SEP vs. non-SEP) without addressing the timing or intensity of SEP events. Future work should explore advanced augmentation, domain adaptation, and real-time integration strategies to improve robustness and operational readiness.

[18] https://gs671-suske.ndc.nasa.gov/

A.18. *UNifying Solar Particle Event modeLLing (UNSPELL) Model*

Model Developers and Relevant Citation: Sigiava Aminalragia-Giamini, Constantinos Papadimitriou, Ingmar Sandberg, Savvas Raptis; Aminalragia-Giamini et al. (2021).

Table 23: Model, Input and Output Specification Table for the UNSPELL model.

Model	
Type	Neural Network
Complexity	81,120*
Input	
Shape	Point Data (0D), Spectra (1D)
Type	Soft X-ray, Flare Location
History	25 years (1988-2013)
Diversity	18,025 samples
Imbalance	0.0128 positive
Sample Size	208 bytes
Sample Coverage	0.4 hours**
Output	
Prediction	Classification, Probability
Type	Triggered
Forecast Window	24 hours
Comments: *UNSPELL is an ensemble of 1000 NNs, each one of which has a complexity of 81120 trainable parameters. **The sample coverage varies, with an average of 25 minutes.	

Summary: A ML approach was introduced in Aminalragia-Giamini et al. (2021) which uses as the main input X-ray measurements data to forecast whether solar flares will lead to SEP events. This model approach and rationale is focused on real-conditions applicability and the production of forecasts during and immediately after the occurrence of solar flares. The model uses publicly and real-time available data which entail GOES X-ray fluxes and the flare detection/definition provided by NOAA SWPC, as well as the flare heliolongitude of the solar flare, when available. The latter is retrieved from *Solar Demon*[19] provided by the Royal Observatory of Belgium. Subsequent work further expanded on the methodology presented in (Aminalragia-Giamini et al. 2021) and lead to the currently operational version integrated in the UNSPELL system as the solar flare module.

Model Description: The model uses a binary classification for SEP and non-SEP occurrence and provides a probability $P \in [0, 1]$ that a flare will lead to an event. The core of the model are NNs which are run in parallel with the same inputs, each providing an independent estimation of the SEP probability occurrence. The final output consists of the mean probability as well as the probability standard deviation resulting from the ensemble outputs. To collapse the probabilistic output to a categorical one, these two outputs are compared against defined thresholds and if both are above or below defined values an alert that an SEP will occur is issued.

Inputs: Model training used 228 samples labeled as positive —flares associated with subsequent SEP events, and 17797 samples labeled as negative— flares not associated with SEP events. The flare X-ray timeseries are used to derive 24 features for a flare which are used as input. If the heliolongitude is also available, its cosine and sine are also used as inputs. The flares used for training are detailed in the NOAA GOES solar flare catalogue and the training dataset spans the years 1988-2013 covering the largest part of SC22, the whole of SC23, and the rising phase of SC24.

Outputs: The output of the model is the ensemble mean probability of SEP occurrence and the ensemble standard deviation of probabilities. The output is converted to a categorical binary label (True/False) using a threshold of 0.7 (above for True) for the average probability, and 0.06 (below for True) for the standard deviation.

Model Configuration: The model that is currently operational is based on the methodology described in Aminalragia-Giamini et al. (2021) with two substantial differences. The NNs used in the development and valida-

[19] https://www.sidc.be/solardemon/

tion were deep networks with several layers and the ensemble members numbers N that were tested were $N = 3$ or $N = 10$. Subsequent investigations showed that ensembles with shallow feed-forward networks with three layers and fewer neurons were able to match the performance previously achieved having the benefit of lower complexity and much faster training times. At the same time the ensemble number N was increased to 1000. While this is a high number, each NN was trained using a different subset of the total available training data, an approach bearing similarities to that of random forests and its variants. This approach was selected in order to achieve good generalization on the ensemble level, avoid any potential overfitting with a singular or few NNs, and have a large enough ensemble from which to derive a meaningful standard deviation of outputs to be used in the subsequent thresholding process.

Model Validation and Results: Our investigation on the operational module reproduced the findings of the original publication for all flares above C1 with true positive rates of 0.86 and true negative rates of 0.92 with a resulting TSS of ~ 0.78. The model has participated in SEPVAL and further validations on the operational outputs will be performed in the near future during the current SC25.

Access to Model Data and Forecasts: The X-ray data and flare list are provided by NOAA at online repositories and the SEP list used is detailed in Pacheco Mateo (2019). The forecasts of the model will be publicly available in 2026 through the ESA *Space Weather Service Network*[20] portal and specifically the *Space Radiation Expert Service Centre*[21], or can contact the authors of Aminalragia-Giamini et al. (2021) directly. For more information follow the links bellow:

- https://www.ncei.noaa.gov/data/goes-space-environment-monitor/access/science/xrs/
- https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/xrs/

Limitations, Caveats and Discussion: A limitation of the model is that it relies solely on solar flare inputs to forecast the SEP occurrence and it does not use CME data. This is an inherent limitation in the model and it was a choice made in its development so that it is able to provide as accurate as possible forecasts, as quickly as possible, since real-time CME data can have large uncertainties and be delayed for several hours. Another limitation is of course the large class imbalance between the positive and negative categories, flares associated and not associated with SEP events. While there is no perfect substitute for real data in the training of ML models, this caveat was addressed to a large degree by taking into account this imbalance in the very training of the model itself.

A.19. *Time Series-Histogram of Oriented Gradients-TaBular (TS-HOG-TB) Model*

Model Developers and Relevant Citation: Pouya Hosseinzadeh, Soukaina Filali Boubrahimi, and Shah Muhammad Hamdi; Hosseinzadeh et al. (2025).

Summary: This study introduces an end-to-end ensemble ML framework for predicting high-impact (~ 100 MeV) SEP events by integrating multimodal data. The proposed model combines three key data modalities: GOES proton flux time series, AR polygons extracted from SOHO EUV images, and solar flare-related tabular data. Each modality is independently evaluated using specialized models, and the best-performing classifiers are combined using ensemble strategies. The final model, the Time Series-Histogram of Oriented Gradients-TaBular (TS-HOG-TB), which integrates all three modalities, achieves strong results with a recall of 0.80 (for balanced) and 0.75 (for imbalanced). The framework shows robustness under noise and across various temporal settings, highlighting the advantage of multimodal fusion for reliable SEP forecasting.

Model Description: The best-performing model, TS-HOG-TB, is an ensemble ML framework that combines predictions from three specialized unimodal models, each trained on a distinct data modality: a) time-series proton flux data from the GOES P6 channel processed using the TSF classifier; b) AR polygon data extracted from SOHO EUV images, encoded with HOG features and classified using random forests; and c) tabular data including sunspot counts, AR counts, and flare class, modeled using an SVM. Each unimodal classifier outputs probabilistic predictions, which are averaged in the ensemble to produce the final classification. This late-fusion strategy captures complementary temporal, spatial, and statistical patterns of solar activity, resulting in improved accuracy and robustness in predicting ~ 100 MeV SEP events.

Inputs: The model takes three distinct data modalities as input: a) GOES time-series proton flux data from the P6 channel (~ 100 MeV), using a 6-hour observation window prior to the associated solar flare; b) EUV images from SOHO's EIT 304 Å channel, processed to extract AR polygons using thresholding, contour detection, and binary

[20] https://swe.ssa.esa.int/

[21] https://swe.ssa.esa.int/space-radiation

Table 24: Model, Input and Output Specification Table for the TS-HOG-TB model.

Model	
Type	Ensemble Method
Complexity	100,000
Input	
Shape	Time Series (1D), Vectors (1D)
Type	EUV Imagery, Energetic Protons
History	15 years (1997-2012)
Diversity	207 samples
Imbalance	0.1449 positive
Sample Size	16.90 bytes
Sample Coverage	6 hours
Output	
Prediction	Classification
Type	Triggered
Forecast Window	0 hours
Comments: *A forecast window of 0 hours indicates that the model performs event-level classification rather than temporal forecasting. These models rely on the GSEP catalog to label whether an SEP event occurs, without predicting its onset time or lead interval. Consequently, the output reflects the presence or absence of an SEP event conditioned on the input observations, not a forward-looking warning horizon.	

masking; c) tabular features including sunspot counts, AR counts, and flare class (C, M, X), computed from the 6-hour period preceding each flare. All SEP events are sourced from the GSEP catalog, while non-SEP events are selected from the HEK database. Events with incomplete data in any modality are excluded.

Outputs: The output of the model is a probability of SEP occurrence based on a flare trigger. The probability is converted to a binary label (SEP/no-SEP), informing us whether a flare will produce an SEP event.

Model Configuration: For time-series input, the model uses TSF, which extracts summary statistics (mean, variance, slope) over random intervals from the 6-hour GOES P6 proton flux data. For image-based input, EUV AR polygons are processed with HOG features and classified using random forest. For tabular input (sunspots, AR counts, flare class), an SVM is used. The final ensemble model (TS-HOG-TB) averages probabilistic outputs from each unimodal classifier. All models are trained with 5-fold cross-validation. Z-score normalization is applied to time-series data; AR images are resized to 200×200 pixels before feature extraction.

Model Validation and Results: Model performance was evaluated using 5-fold cross-validation under both balanced and imbalanced settings. For the balanced setting (37 SEP, 37 non-SEP), the TS-HOG-TB ensemble achieved 0.81 recall, 0.80 F1-score, and the highest TSS and HSS scores among all models. In the imbalanced setting (37 SEP, 104 non-SEP), the model maintained strong Recall (0.75), indicating robustness under real-world class imbalance. Sensitivity analysis with Gaussian noise (mean=0, std=0.5) confirmed that TS-HOG-TB outperforms unimodal models (TS-only, AR-only) under noisy conditions. The model also showed stable performance with reduced training data and achieved best results with a 6–7 hour observation window.

Access to Model Data and Forecasts: The implementation of the TS-HOG-TB ensemble model, including preprocessing scripts and training configurations, is publicly available at https://github.com/pouyahosseinzadeh/Solar-Energetic-Particle-Event-Prediction-Ensemble-TS-HOG-TB. All input data modalities used in this study are obtained from publicly accessible heliophysics repositories. Time-series proton flux measurements are retrieved from the GOES SEM archives provided by NOAA NCEI, EUV AR imagery is accessed via the SOHO/EIT instrument through the *Helioviewer* platform, and tabular solar activity parameters (e.g., sunspot counts, AR counts, flare class) are obtained from the HEK. These resources enable reproducible multimodal SEP forecasting experiments using combined temporal, spatial, and tabular predictors.

Limitations, Caveats and Discussion: A primary limitation is the small number of high-energy ($\sim$ 100 MeV) SEP events with complete multimodal data, which restricts dataset size. While the model performs well in balanced settings, performance varies under class imbalance. Non-SEP event selection required manual filtering to avoid temporal overlap, limiting scalability. EUV image resolution and timing may also introduce uncertainty, as AR segmentation depends on pre-flare image availability. Despite these constraints, the ensemble model demonstrated robustness to noise and reduced training data. Future work will address real-time prediction, expand the non-SEP set, and explore physics-informed ensemble methods.

A.20. *Solar Energetic Particle Network (SEPNET) Model*

Model Developers and Relevant Citation: Yian Yu, Yang Chen, Lulu Zhao, Kathryn Whitman, Ward Manchester, Tamas Gombosi; Yu et al. (2025).

Table 25: Model, Input and Output Specification Table for the SEPNET model.

Model	
Type	Neural Network
Complexity	130,000
Input	
Shape	Time Series (1D)
Type	Magnetic Fields, Soft X-ray
History	40 years (1986-2025)
Diversity	11,773 samples
Imbalance	0.3004 positive
Sample Size	105000 bytes
Sample Coverage	24 hours
Output	
Prediction	Classification, Probability
Type	Continuous
Forecast Window	24 hours

Summary: We introduce SEPNET (and its extensions, SEPNET-TS and SEPNET-O), an innovative multi-task NN that integrates forecasting of solar flares and CME summary statistics into the SEP prediction model, leveraging their shared dependence on SHARP magnetic field parameters.

Model Description: SEPNET incorporates long short-term memory and transformer architectures to capture contextual dependencies in temporally evolving features for SEP forecasting.

Inputs: For each sample, the input consists of a set of min-max normalized features derived from solar flare, CME, and SHARP magnetic field data.

Outputs: The SEPNET model, together with SEPNET-TS, is trained using all SEP event enhancements above GOES background, indicated by a proton flux threshold of 10^{-6} pfu in the CLEAR SEP benchmark dataset. For operational deployment (SEPNET-O), samples labeled as operational SEP events ($\geq$ 10 MeV proton flux $\geq$ 10 pfu) are used as a validation set to fine-tune the classification threshold for distinguishing operational SEP events.

Model Configuration: The input features are processed through three shared fully connected (dense) layers with gradually reduced feature dimensionality (from 256 to 128, 64, and 16). Each dense layer is followed by layer normalization, ReLU activation, and dropout. The shared embedding is then fed into two distinct output heads to implement multi-task learning: a regression head that predicts the counts of future flare and CME events, and a classification head that outputs the predicted probability of a future SEP event. To better capture temporal dependencies and complex sequential patterns in the input data, the updated model `SEPNET-TS` integrates recurrent and attention mechanisms by combining a unidirectional LSTM layer with a transformer encoder.

Model Validation and Results: The performance of SEPNET is evaluated on the state-of-the-art SEPVAL SEP dataset and compared with classical ML methods and current state-of-the-art pre-eruptive SEP prediction models. The results show that SEPNET achieves higher detection rates and skill scores while being suitable for real-time space weather alert operations.

Access to Model Data and Forecasts: The operational forecasting is available in real-time (updated every hour) through the MLSW website at https://mlsw.engin.umich.edu/apps/runSEP. The code and data are available at: https://github.com/yuyian/SEP-Prediction.git.

Limitations, Caveats and Discussion: In our follow-up ongoing work, we extend the forecasting to incorporate X-ray and proton flux history, in predicting both SEP occurrence and corresponding proton flux values.

A.21. *Bidirectional Long Short-Term Memory (BiLSTM-SEP) Model*

Model Developers and Relevant Citation: Mohamed Nedal, Kamen Kozarev, Nestor Arsenov, and Peijin Zhang; Nedal et al. (2023).

Table 26: Model, Input and Output Specification Table for the BiLSTM-SEP model.

Model	
Type	Neural Network
Complexity	333,699
Input	
Shape	Time Series (1D)
Type	Magnetic Fields, Soft X-ray, Proton Flux, Ground-Based Radio, Solar Wind
History	43 years (43)
Diversity	15,558 samples*
Imbalance	0.0146 positive
Sample Size	15,120 bytes
Sample Coverage	6,480 hours**
Output	
Prediction	Classification (All Clear), Probability, Regression
Type	Continuous
Forecast Window	24 hours***
Comments: * BiLSTM-SEP does time series forecast therefore diversity reflects the number of days in the training samples. The total number of days is 15,558, from December 25th 1976 to July 30th 2019. ** For every sample, time-series history of 270 days (6480 hours) is used. *** A forecast of 24 hours ahead produces the best results. Forecast windows of 48 and 72 hours (2 and 3 days) are explored too.	

Summary: The Bidirectional Long Short-Term Memory (BiLSTM-SEP) is a deep learning model for forecasting SEP fluxes across three GOES energy channels using bidirectional LSTM networks. The model processes multi-decadal, multivariate time series of solar and heliospheric data to generate 3-day forecasts of log-integral proton flux. Designed to capture both short and long-term dependencies in space weather data, it performs competitively with other forecasting approaches while maintaining low FAR. Its outputs are suited for both operational radiation hazard mitigation and downstream scientific analysis.

Model Description: The model consists of four BiLSTM layers with 64 neurons each, followed by a dense layer that outputs a 3-day sequence of predictions. Its bidirectional architecture allows the model to learn from both past and forward temporal context, which improves performance on highly nonlinear and non-stationary SEP data. Each GOES energy channel is modeled independently. Training uses early stopping, adaptive learning rate reduction, and the Huber loss function (Meyer 2021) for robustness to outliers.

Inputs: Seven input features were chosen for their physical relevance and correlation with SEP flux: sunspot number, F10.7 index, long and short-band X-ray fluxes (log-transformed), solar wind speed, interplanetary magnetic field magnitude, and prior log-SEP fluxes in each energy band. All features were daily averaged, normalized, and linearly interpolated to fill gaps. The model uses a sliding window of 270 days of inputs to forecast the next 3 days.

Outputs: Each trained model predicts the next three days of log-integral proton flux for a specific energy channel ($\geq$ 10 MeV, $\geq$ 30 MeV, or $\geq$ 60 MeV). Outputs are real-valued (not categorical) and can be used in radiation exposure

calculations or space weather alert systems. The multi-input ulti-output strategy avoids recursive input-feedback and allows efficient, simultaneous multi-step prediction.

Model Configuration: The model was trained using a batch size of 30 days (approximately a Carrington rotation), with an input window of 270 days. A training set (74.29%), validation set (16.2%), and test set (9.51%) were carved out using a fixed 9:2:1 month-based strategy from each year. The Adam optimizer and Huber loss were used, with learning rate reduction and early stopping. Separate models were trained per energy band to reduce cross-channel interference.

Model Validation and Results: On validation and test sets, the model achieved $R \geq 0.9$ for all energy bands at a 1-day lead time. MAE ranged from 0.045 to 0.125 across channels and lead times. Mean Absolute Percentage Error ranged from 12.36 to 49.14. Performance declined slightly with increasing forecast horizon, as expected. The ≥ 60 MeV model showed the highest consistency, while the ≥ 30 MeV model showed a somewhat larger discrepancy between predictions and observations. Confusion-matrix-based skill scores showed a low FAR and competitive POD compared to prior models such as UMASEP Núñez (2011) and Relativistic Electron Alert System for Exploration (REleASE; Malandraki & Crosby 2018).

Access to Model Data and Forecasts: Data and preprocessing scripts are available at https://gitlab.com/iahelio/mosaiics/sep-lstm/. All data used are publicly available from OMNIWeb at https://omniweb.gsfc.nasa.gov, GOES SEM archives at https://satdat.ngdc.noaa.gov/sem/goes/data/avg, and Solar Influences Data Analysis Center (SILSO; sunspot number) at https://www.sidc.be/silso/home. Forecast near-real-time model outputs are under development and will be made available in a future public repository, as part of an effort to support operational space weather forecasting.

Limitations, Caveats and Discussion: While overall results are encouraging, the model still struggles to forecast rare high-flux events (e.g., SEP ≥ 10 pfu) due to the inherent data imbalance —these events are underrepresented in the dataset, often leading to underestimation of their peak flux. Additionally, forecasts for the ≥ 30 MeV energy channel exhibited more pronounced deviations from observations than the ≥ 10 and ≥ 60 MeV models. Forecast performance also decreased during solar minimum and quiet periods, where the predictive signal in inputs is weak. Addressing these limitations will require higher-resolution (e.g., hourly) inputs, more sophisticated feature engineering, and additional data from SC 25.

A.22. *Models for Probabilistic Forecast of Solar Energetic Particles (MEMPSEP)*

Model Developers and Relevant Citation: Subhamoy Chatterjee, Maher Dayeh, Andrés Muñoz-Jaramillo, Kim Moreland, Hazel Bain, Samuel Hart, Michael Starkey; Chatterjee et al. (2024), Dayeh et al. (2024) and Moreland et al. (2024).

Summary: The Multivariate Ensemble of Models for Probabilistic Forecast of Solar Energetic Particles (MEMPSEP; Figure 7) introduced an ensemble of ML models to predict if a solar flare would lead to SEP events and, if so, what the properties of that event would be with associated uncertainty. The model ingests both remote-sensing (SoHO/MDI and SDO/HMI) and in-situ data over 1998-2013 to make predictions.

Model Description: A CNN architecture was built that ingests multiple inputs such as images, scalar parameters, and time series to make SEP event/non-event classification and regression of SEP properties. The CNN processes the input image sequences with repeated 2D convolution, batch-normalization, non-linear activation, and max-pooling. The flattened layers are concatenated with scalar parameters and features extracted from time series data through 1D convolution. The concatenated features are passed through dense (fully connected) layers to reproduce the classification and regression targets. First, the classification model is trained, followed by the regression model in two different ways: gated and non-gated. For the gated approach, the regression is trained by coupling the classification model outcome to the regression loss function, and for the non-gated approach, the regression model is trained independently.

Inputs: The GOES integrated (≥ 10 MeV) particle flux is checked for whether it crosses a threshold of 5 pfu within 6 hours of flare onset. If yes, it is then labeled as an SEP event and otherwise as a non-event. The inputs to the model are acquired over 1-3 days prior to the flare onset. A 3-day preflare magnetogram time sequence is used as remote sensing input. For in-situ, WIND/WAVES radio burst time-freq map over 3 days, X-ray, and L1 Electron time series over 1 day before flare, and scalar properties related to solar wind and suprathermal population are used. A total of 10200 non-events and 675 events for training the model ensemble are utilized.

Outputs: MEPSEP produces SEP occurrence probability and SEP properties such as proton peak fluxes (≥ 5, ≥ 10, ≥ 30, ≥ 60, ≥ 100 MeV), event onset, and duration.

Table 27: Model, Input and Output Specification Table for the MEMPSEP model.

Model	
Type	CNN
Complexity	6,092,617
Input	
Shape	Point data (0D), Time-series (1D), Images (2D)*
Type	Magnetic Field, Soft X-ray, Electron Flux, EUV Imagery, Coronagraphs, Space-Based Radio, Solar Wind
History	15 years (1998–2013)
Diversity	13,850 samples
Imbalance	0.07 positive
Sample Size	17,000,000 bytes**
Sample Coverage	72 hours***
Output	
Prediction	Classification, Probability, Regression (Time-series)
Type	Triggered
Forecast Window	6 hours
Comments: * Time Series of magnetograms (3D), Time Series of X-ray flux(1D), Time series of Electron (1D), Time-frequency map of Radio bursts (2D), Scalar solar wind and suprathermal particle properties (0D) ** For each datapoint-1.7 MB for magnetograms, 272KB for wind/waves time-freq map, 60 KB for X-ray time-series, 500 KB for electron time-series *** 3 days for magnetogram sequence and Wind/Waves time-frequency map prior to flare onset, 1 day for X-ray and electron time series.	

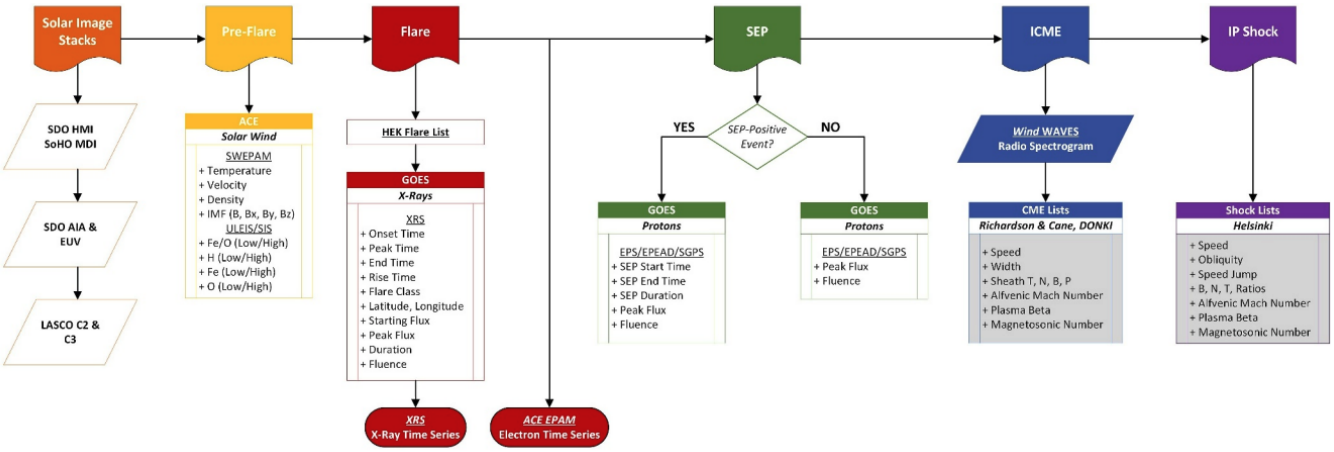


Figure 7: The complete data set flowchart shows all incorporated observations from remote imaging to in-situ measurements. Each section relates to a time frame in the event series: Solar images (pre-flare, flare), solar wind conditions (pre-flare), X-ray properties, and time series (flare). Post-flare properties for the SEP event, interplanetary CME parameters with WIND/WAVES radio spectrogram, interplanetary shock properties, and published shock lists. We note the observation (italics) and the instrumentation (underlined) used to obtain in-situ data along with each parameter observed or calculated parameter.

Model Configuration: A large class imbalance of SEP events and non-events poses difficulty in the success of ML-based prediction models. We made use of the class imbalance, creating multiple training and validation sets through stratified undersampling of non-events and the same set of events. With each of those training and validation sets, a CNN is trained with the architecture described previously. The sets generated an ensemble of models for both the classification and regression tasks. The CNN classifier is turned into a predictor of true probability using a

non-parametric calibration technique called Bayesian Binning Quantile. An ensemble of BBQ calibrators was derived, providing uncertainty estimates on the SEP occurrence probability. Among the non-gated and gated regression models, MEMPSEP achieved better results with the gated model.

Model Validation and Results: A test set is designed including events/non-events that are not seen by the model during the training phase for estimating both probabilistic and deterministic skill scores. A Brier Score (BS) of 0.14 is achieved and an Expected Calibration Error (ECE) of 0.07 on the ensemble median probability. Applying a threshold of 0.5, a POD, False Positive Rate (FPR), TSS, and HSS of 0.83, 0.2, 0.63, and 0.6, respectively is achieved. For the regression model, an R^2 score of ≥ 0.63, 0.07, and 0.01 for SEP peak, onset, and duration with occurrence probability ≥ 0.5 is achieved. MEMPSEP participated in the Solar Heliospheric and INterplanetary Environment (SHINE) 2022 independent validation as well as SEPVAL campaign, providing a BS of 0.2, and predicting 6 out of 8 events and 11 out of 14 non-events correctly with a probability threshold of 0.5.

Access to Model Data and Forecasts: The MEMPSEP training data and codes have been made publicly available. They can be accessed via the following links: https://zenodo.org/records/10044865 and https://doi.org/10.5281/zenodo.11201195.

Limitations, Caveats and Discussion: The model ensemble is currently flare-triggered and cannot forecast the full temporal profile of the SEP fluxes. Although the model currently ingests Magnetograms as imaging data, the MEMPSEP dataset defined in Moreland et al. (2024) also consists of EUV and coronagraph imagery. The impact of those additional inputs on MEMPSEP performance is being tested. We are currently developing MEMPSEP further to eliminate the flare trigger requirement and enable rolling predictions of time series through the course of an SEP event using near-real-time data (Dayeh et al. 2025). An effort is also ongoing to add the energetic storm particle sudden enhancements, often associated with SEPs.

A.23. *Parker Solar Probe SEP Prediction (PSPSP) Model*

Model Developers and Relevant Citation: Tate Hutchins, Spiridon Kasapis, Hameedullah Farooki, Manuel Cuesta, Lengying Khoo2, Sungmin Pak, Robert Czarnota, Jamie Rankin, Jamey Szelay, Georgios Livadiotis, Xiaoyan Li, David McComas, Zigong Xu, Nikolaos Sarlis; Hutchins & Kasapis (2026).

Table 28: Model, Input and Output Specification Table for the PSPSP Model.

Model	
Type	Neural Network
Complexity	13,814,081
Input	
Shape	Time Series (1D), Images (2D)
Type	EUV Imagery, Proton Flux, Solar Wind
History	6 years (2019-2025)
Diversity	1,015 samples
Imbalance	0.1547 positive
Sample Size	6,300,000 bytes
Sample Coverage	240 hours
Output	
Prediction	Classification, Regression
Type	Continuous
Forecast Window	35 hours
Comments: *The forecast window varies from 2 hours to 3 days, with a median value of 35 hours.	

Summary: Most ML models in this document aim to predict geoeffective SEP events. The PSPSP model uses EUV images from the AIA onboard the SDO to predict the intensity values within any given point in ecliptic plane between Sun and Earth. To train the model, PSP Integrated Science Investigation of the Sun (IS⊙IS; McComas et al. 2016) Energetic Particle Instruments - Low (EPI-Lo; Hill et al. 2017; **?**) particle intensity measurements were used as targets. A series of convolutional layers extract image features, which are combined with information about the

location of PSP and passed through a set of dense layers in order to provide a binary classification about whether the particle population in the given space is ≥ 0.1.

Model Description: The study focuses on a snapshot model focusing on single image input and a better performing video model focusing on a chronological sequence of images input. Each model consists of a configurable-depth CNN that extracts spatial features from the AIA 171$\mathring{A}$ inputs using successive Convolution, Batch Normalization, ReLU and Maximum Pooling blocks. Global average pooling is applied to obtain a fixed-length feature vector, which is concatenated with the PSP magnetic footpoint information. For the video model, a transformer encoder layer learns and appends a non-local temporal embedding to learn the progression of SEP events. The combined representation is passed through a fully connected MLP with ReLU activations and dropout regularization to produce a single output prediction of intensity.

Inputs: The inputs to the model are a) SDO AIA 171$\mathring{A}$ images that are reduced in dimension (512×512 pixels) and b) the solar footprint longitude and distance values taken from the PSP ephemeris data. The solar footpoint longitude is calculating assuming the Parker spiral utilizing the ephermeris data and the PSP solar-wind speed, which is derived from measurements of the Solar Wind Electrons Alphas and Protons (SWEAP; Kasper et al. 2016) instrument. Here, we use the level 3 Solar Probe Analyzer for Ions (SPAN-I; Livi et al. 2022) solar wind proton speed magnitude. As targets to the model we use the energy weighted average (J_{linlin}, Cuesta et al. 2025) particle intensity measured by the PSP IS$\odot$IS EPI-Lo instrument. The inputs span 6 years, from the beginning of the PSP mission to December 2025. Both inputs are log-normalized and then min-max normalized too.

Outputs: The model outputs a prediction for the J_{linlin} particle intensity on the given point of the PSP orbit. To provide SEP event classification predictions, a threshold is set at 10^{-1} such that any J_{linlin} reading above is defined as a positive event and anything else is defined as a negative non-event. Models were trained on targets of the actual J_{linlin} values as well as modified and trained on the 0/1 classification values.

Model Configuration: The best performing model has a batch size of 32 images, dropout rate of 0.25, 4 attention heads, 3 attention blocks, hidden head size of 256, and image embedding size of 256. It was trained using a learning rate of 0.0001 and achieved the best loss after 31 epochs.

Model Validation and Results: With a 0.1 J_{linlin} threshold, the Video model trained on classification targets achieves an Accuracy of 0.7705, Precision of 0.3803, Recall of 0.8039, FAR of 0.2355, TSS of 0.5684, and HSS of 0.3902.

Access to Model Data and Forecasts: All relevant code can be found at https://github.com/thutch17/PSP-SEP-Event-Prediction.

Limitations, Caveats and Discussion: The main limitation of this study is that the model does not effectively capture temporal relationships, considering the video model only performed marginally better than the snapshot model. Future studies should use model architectures that can capture temporal dependencies not only in the J_{linlin} timelines but also in the solar disc progression captured in the SDO AIA image series.

A.24. *Energetic Particle Radiation Environment Module - S (EPREM-S) Model*

Model Developers and Relevant Citation: Atilum Gunes Baydin, Bala Poduval; Baydin et al. (2023).

Summary: The challenge in developing an ML model for the prediction of SEPs is the lack of sufficient number of SEP events for training and validation of the ML model —the class imbalance problem— despite decades of SEP observations by spacecraft. Availability of synthetic (or simulated) SEP events created using first principles models will solve the class imbalance problem to a large extent. However, many first principles models are computationally intensive and takes tens of minutes to hours for completing one simulation. Therefore, simulating hundreds of thousands of SEP events for training the ML model becomes rather impractical within a reasonable time-frame. This difficulty can be overcome with the method of emulation or surrogate models where a NN model trained on the simulated out put (SEPs, for example) of a first principle model is developed to perform the exact same function as that of the original model with acceptable accuracy but much faster. Baydin et al. (2023) describes the method and the results of the NN surrogate model of the Energetic Particle Radiation Environment Model (EPREM) developed by Schwadron et al. (2010). For this, 32,000 SEP events are generated to train a feed-forward NN, EPREM-S (Figure 8), with 4 hidden layers and ReLU nonlinearities after each layer except the last one. It should be noted that the EPREM and EPREM-S outputs are in remarkable agreement, the MSE being 0.07 as it is found during validation of the surrogate model. Analysis of an event previously unseen by the surrogate model as a Bayesian inference problem revealed that the parameters were correctly inferred as their ground truth values (unknown to the inference algorithm) are contained

Table 29: Model, Input and Output Specification Table for the Baydin et al. (2023) Model.

Model	
Type	Neural Networks
Complexity	285,881,344
Input	
Shape	Time Series (1D)
Type	Proton Flux
History	N/A*
Diversity	32000 samples
Imbalance	1**
Sample Size	1,600,00 bytes
Sample Coverage	96
Output	
Prediction	N/A***
Type	Continuous
Forecast Window	N/A***
Comments: *This study does not use observational data but rather synthetic (simulated) data that contained 32,000 SEP events. **Only positive events are used in this simulation study. ***This is a surrogate model capable of SEP prediction but no prediction accuracy analysis carried out.	

within the resulting posterior distributions. By measuring the runtime costs of EPREM and EPREM-S, it is found that EPREM-S is tens to hundreds of thousands times (10^4 - 10^6) faster than EPREM.

Model Description: The results are based on a feed-forward NN with four hidden layers of sizes 512, 1,024, 2,048, and 138,240, and ReLU nonlinearities after each layer except the last one, giving rise to a total number of 285,881,344 trainable parameters. The output of the last layer is reshaped into a cube with shape $24 \times 288 \times 20$ (24 streams, 288 time steps, and 20 energy levels). In order to provide uncertainty quantification when running trained EPREM-S models, a deep ensemble (Lakshminarayanan et al. 2017) approach is used, where multiple independently trained EPREM-S instances are involved. Following standard practice, these model instances are trained with the same data, but using a different random number seed leading to different model weight initialization and course of stochastic optimization for each instance. Given the set of pretrained surrogate models $S_i = 1, 2, ..., M$, and a new event parameter ψ, the mean and standard deviation of the flux predictions were estimated as

$$\mu_\phi = \frac{1}{M} \Sigma S_i(\psi) \tag{A1}$$

$$\sigma_\phi = \sqrt{\frac{1}{M} \Sigma S_i(\psi) - \mu_\phi} \tag{A2}$$

Inputs: EPREM simulates the SEPs by solving the focused transport equation numerically in a Lagrangian frame of reference (Schwadron et al. 2010). Nested cubes the surfaces of which are subdivided into square cells with the grid nodes at their centers make up the EPREM grid. The grids at the inner boundary corotates with the Sun at each time step and the nested shells advance radially outward with the solar wind. EPREM will compute the distribution function of a given source of particles such as SEPs or pickup ions, anywhere in the heliosphere at individual nodes advecting with the speed of solar wind, naturally tracing the Parker spirals. For this, EPREM makes use of a seed particle spectrum that is a function of energy and heliocentric distance. Five core parameters of the initial seed function are selected, namely, boundary function amplitude, energy spectrum power-law index γ, radial scaling index β, boundary function cut-off energy, also called roll over energy or knee cut-off, and the mean free path, as the variable input parameters, that is, the parameters which EPREM-S takes as input.

Outputs: The model output is time series of particle fluxes as a function of time, heliocentric distance and energy. The simulation domain consists of 4 days and there are 20 energy levels from 0.01 - 200 MeV.

Model Configuration: A feed forward NN with four layers and about 286 million learnable parameters is used.

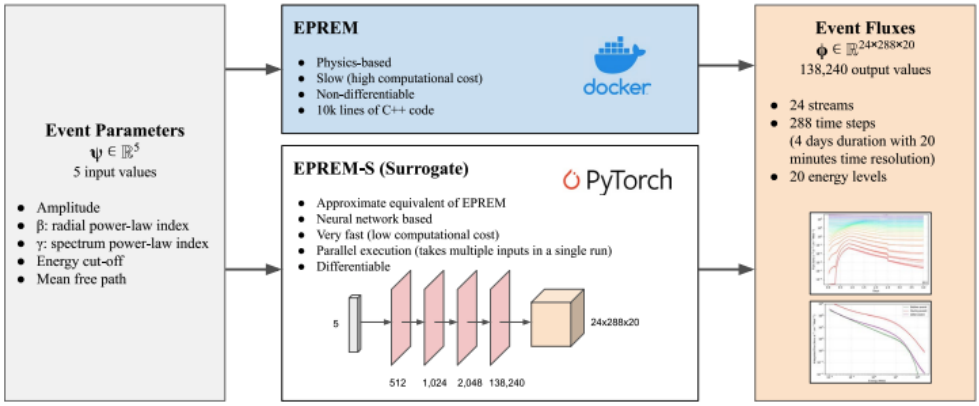


Figure 8: Overview of EPREM and EPREM-S models. Both map input parameters ψ to output fluxes ϕ. EPREM is a physics-based model for generating SEP events implemented in C++. EPREM-S is a NN trained with a data set $D = \{\psi_i, \phi_i = EPREM(\psi_i)\}_{i=1}^{N}$ obtained by running EPREM with inputs ψ_i $p(\psi)$ sampled from a continuous uniform prior distribution.

Model Validation and Results: Surrogate modeling or emulation is the creation of fast and simple models that approximate the behavior of complex analytical models that are computationally expensive to evaluate (Queipo et al. 2005; Forrester et al. 2008). An ensemble of five independently trained EPREM-S models is trained, where each model is a feed-forward NN with four layers and approximately 286 million learnable parameters. Using the EPREM-S ensemble, several thousand SEP events are generated for a range of values for five physical parameters of the initial seed spectrum in EPREM. A comparison of the outputs of EPREM and EPREM-S showed remarkable agreement supported by the fact that the MSE was 0.07 during validation of the surrogate model. EPREM-S was used for simulation-based inference by treating the range of values for the selected five parameters of the initial seed spectrum as the prior distribution for inferring the posterior distribution. This has been demonstrated using a set of events unseen by EPREM-S during training where in all the cases the ground truth values of the selected events were found to be contained within the posterior distributions.

Access to Model Data and Forecasts: The 32,000 SEP events simulated using EPREM and the codes are available at https://zenodo.org/records/10109868, and https://zenodo.org/records/10038847.

Limitations, Caveats and Discussion: EPREM-S is a generative model which can be used to predict SEP events and also for simulation-based inference of observed events. Since EPREM-S is designed and trained on five specific parameters of the initial seed spectrum, the inferences will primarily be based on the influence of the seed population in the time profiles and evolution of SEP events. Surrogates trained on more physical parameters of the first principle model EPREM, giving rise to more general-purpose surrogates, will be the continuation of the emulation work presented here.

B. DESCRIPTION OF DATASETS FOR SEP PREDICTION

B.1. *MEMPSEP-III Dataset*

The MEMPSEP-III dataset (Moreland et al. 2024) is an ML-oriented multivariate dataset specifically designed for forecasting the occurrence and properties of SEP events. It integrates both in-situ and remote sensing observations from multiple spacecraft, including GOES, ACE, SDO, SOHO, and WIND/WAVES, covering SC 23 and part of SC 24 (1998–2013). The dataset comprises 252 SEP-producing solar flare events and 17,542 non-SEP events, identified using the GOES flare event list. For each event, MEMPSEP-III includes a rich set of features such as energetic proton and electron fluxes, upstream solar wind parameters, interplanetary magnetic field vectors, and remote solar imaging and radio observations. This multivariate structure enables flexible input configurations for ML models and supports both classification and regression tasks related to SEP forecasting. The MEMPSEP-III dataset has been carefully curated, validated, and cleaned to ensure reliability for ML applications. It has been used to train the MEMPSEP, as described

in a series of accompanying papers (Chatterjee et al. 2024; Dayeh et al. 2024). Its design facilitates experimentation with different feature sets and model architectures, making it a valuable resource for advancing data-driven SEP prediction.

B.2. *MTS-SEP Dataset*

Hosseinzadeh et al. (2024a) present a dataset and methodology aimed at improving the prediction of high-energy SEP events, particularly those involving ~ 30, ~ 60, and ~ 100 MeV protons. A key challenge addressed in this work is the scarcity of SEP events, which limits the effectiveness of ML models. To overcome this, the authors apply data augmentation techniques to synthetically increase the number of SEP samples, thereby enhancing model performance. The dataset consists of univariate and multivariate time series of proton flux measurements, spanning SCs 22 to 24. These time series are used as input to ML classifiers, with a particular focus on the TSF algorithm. The authors demonstrate that using multivariate time series data significantly improves prediction accuracy, especially for the ~ 100 MeV SEP events. By applying the SMOTE, they report a 20% increase in average accuracy, reaching approximately 90% for the highest energy SEP prediction task. In addition to the dataset, the authors develop a pipeline framework for hierarchical classification of SEP and non-SEP events across different energy thresholds. This work highlights the importance of data augmentation and multivariate temporal features in enhancing the predictive capabilities of ML models for SEP forecasting.

B.3. *GSEP Dataset*

Rotti et al. (2022b) introduce the integrated Geostationary Solar Energetic Particle Events Catalog (GSEP), a homogenized dataset of SEP events spanning SCs 22 to 24. The catalog is constructed by correlating and integrating three existing SEP datasets based on GOES integral proton flux measurements. Each event in the catalog has been visually verified and labeled to ensure consistency and reliability. It has been revised in Rotti & Martens (2023) to include additional weak SEP events. The latest GSEP catalog identifies a total of 433 SEP events, of which 244 exceed the SWPC threshold for significant proton events. For each event, the dataset includes sliced time-series data of proton flux intensity profiles across multiple energy bands, along with metadata describing associated solar eruptions such as flares and CMEs. This dataset is publicly available and designed to support ML and statistical analyses of SEP events and their solar sources. Its structured format and validated event labeling make it a valuable resource for developing predictive models and improving space weather forecasting capabilities.

B.4. *SMARP-SHARP Dataset*

Kosovich et al. (2024) present a merged dataset of magnetic field parameters derived from two major solar AR data products: the SMARPs (Bobra et al. 2021) from the MDI instrument onboard SOHO and the SHARPs (Bobra et al. 2014) from the HMI instrument onboard SDO. This unified dataset spans from 1996 April 4 to 2022 December 13 and is designed to support solar flare and SEP forecasting and broader space weather applications. The merging process involves filtering, rescaling, and combining SMARP and SHARP parameters into uniform multivariate time series representations of solar ARs. These time series can be spatially reduced and correlated with other space weather indicators, such as the daily solar flare index and soft X-ray flux measurements from GOES satellites. Preliminary statistical analysis using time-lagged cross-correlation and rolling-window techniques reveal that certain magnetic field properties of ARs may precede flare activity, suggesting potential predictive relationships. The dataset enables exploration of these dynamics across multiple solar cycles and provides a foundation for developing ML models that incorporate magnetic field evolution as a predictive feature for solar flares and SEP events.

B.5. *CLEAR Dataset*

The Center for All-Clear Solar Energetic Particle Forecasts (CLEAR) Space Weather Center of Excellence[22] has developed the CLEAR SEP Benchmark Dataset derived from GOES data between January 1986 and September 2025. This dataset emphasizes consistent and automated identification of SEP enhancements using the `fetchsep`[23] tool. The proton time series for all GOES satellites from GOES-06 to GOES-18 were independently analyzed with `fetchsep` to calculate mean background levels and identify all SEP enhancements by applying the event definitions

[22] https://ccmc.gsfc.nasa.gov/swxcoe/

[23] https://github.com/ktindiana/fetchsep

specified in Table 30. The full benchmark dataset package (∼ 30 GB) consists of SEP and non-event (quiet-time period) lists for each GOES satellite, along with the complete time series of the original GOES fluxes, calculated mean background, plots, and other information. The final SEP Benchmark List (∼ 2 MB) was compiled by selecting SEP event information from the primary GOES satellite at the time of the event. The benchmark list includes associated flare, CME, radio, and solar wind information extracted from SEP lists maintained by A. Steve Johnson (NASA SRAG) and Ian Richardson (University of Maryland, NASA GSFC) spanning SCs 22 to 25. Two sub-lists are provided in the Benchmark dataset —the Operational List and the Energy-Bin Calibrated List. The Operational List (1986-2025) uses the archived GOES integral fluxes without background subtraction or modification. The proton values in this list represent data streams used by operational end-users for decision-making. The Energy-Bin Calibrated list (2010-2017) is derived from GOES-13 and GOES-15 background-subtracted uncorrected differential fluxes, with calibrated energy bins provided by Sandberg et al. (2014) and Bruno (2017). Integral fluxes were then estimated from the differential channels. The calibrated list provides a better representation of the SEP energy spectrum. The CLEAR SEP Benchmark Dataset is hosted by CCMC and may be downloaded from https://ccmc.gsfc.nasa.gov/swxcoe/clear/. To promote transparency, reproducibility, and continued maintenance of this dataset, the `fetchsep` repository includes scripts that may be used to generate the benchmark dataset from scratch.

Table 30: Number of SEP events in the CLEAR Benchmark Dataset

Event Definition	Operational List	Energy-Bin Calibrated List
$\geq$ 10 MeV above background	565	98
$\geq$ 10 MeV $\geq$ 10 pfu	265	47
$\geq$ 30 MeV above background	358	43
$\geq$ 30 MeV $\geq$ 1 pfu	258	33
$\geq$ 50 MeV above background	292	24
$\geq$ 50 MeV $\geq$ 1 pfu	161	22
$\geq$ 100 MeV above background	163	24
$\geq$ 100 MeV $\geq$ 1 pfu	88	8

C. QUESTIONNAIRE

Below we present the ML model taxonomy form that was filled out by the SEP modelers of the 23 different publications included in this review.

ML Models for SEP Prediction - Taxonomy Form

Dear co-authors, thank you once again for agreeing to participate in our effort to put together a review paper for the ML models that predict SEP events. As a first step of our collaboration, I would like to ask you to fill out the following form which will be used for comparing the different models in the manuscript and will also help us map the research field of SEP prediction using ML.

The form contains three parts: *Architecture*, *Input* and *Output* Information. Every section has two different types of questions/classifications: quantifiable (you should provide a number) and categorical (you should provide text/choose options). Each section provides more detailed information about the questions asked. Please take your time to fill out the form. Many questions will require some research from your side in order to answer. Please try to answer all questions to the best of your ability.

If your work/published manuscript includes more than one model (for example works that test multiple different ML models), please provide information for the best performing model. If you have trained models that are substantially different from each other and therefore you would like to submit more than one pages for the review manuscript, please submit this form again, one time for each model. This also applies to cases where the authors have published different manuscripts for each model.

Model Information

In this Section we would like you to provide information about the ML model you have trained.

- **Model Type - ML Class:**
  What is the type/architecture of your model? Some examples would be SVMs, Regression Models, Deep NNs,

Random Forests, LSTM architectures etc.
Categorical Response: Open.

- **Model Complexity - Number of Trainable Parameters (n):**
  How deep is your model? Here you need to answer with the number (scalar) of free parameters that your model includes. Do not confuse this number with the number of hyperparameters (parameters chosen by user before training), here we are looking for the number of trainable weights included within your model.
  Quantifiable Response: Open.

Input Information

Here we would like you to provide information about the inputs that you have used to train the model. Most questions are related to the training input, but there are also some questions that are relevant to the model inputs during validation.

- **Shape of Input Data:**
  Here we would like you to select what is the shape (1D/2D/3D) of your input data. Note that here we are looking for the shape of a single occurrence/event, i.e. if your model is trained on multiple time series events, therefore constructing a 2D input matrix, the shape of your input data is still 1D (time series rather than a matrix). You can select multiple choices in the case where you train your model with data of multiple sizes (ex. you input during training both time series and image data to the model).
  Categorical Response: Point Data (0D Features), Time Series (1D), Images (2D or 3D), Spectra (1D), Other.
- **Type of Data - Physical Quantity:**
  Here we would like you to select the physical quantity/ies that your input training data represents. For more information on the input type categories please check Whitman et al. (2023) (Table 10).
  Categorical Response: UV / EUV Imagery, Magnetic Fields / Magnetograms, Electric Fields, X-Ray / Soft X-Ray Intensity, White Light / Optical Imaging, Ground-Based Radio, Space-Based Radio, Coronagraph, Solar Wind (n, T, p, V), Suprathermal Particles, Energetic Protons, Other.
- **Input History - Time Coverage of Training Set:**
  Here we would like you to type in the number of months/years your training data covers. For example, if your very first training event is in 2010 and the last one occurred in 2020, this would be 10 years worth of data. Some studies use multiple solar cycles worth of data, therefore the answer here would be a value greater than 11 years.
  Quantifiable Response: Open.
- **Input Diversity - Number of Events - Total Training Samples:**
  Here we would like to know the total amount of events you use as a training input (or targets in some applications). For example, if you predict SEPs based on solar flare occurrences, we would like to know the total number of positive (SEP producing) and negative (non-SEP producing) flare events you have used. For many studies this can be simply the total number of SEP events used as targets. Please do not take in account here the number of events you use to validate your model. Positive Samples + Negative Samples = Total Training Samples
  Quantifiable Response: Open.
- **Class Imbalance - Percentage of Positive Samples in Training Set:**
  Here we need you to answer with a number between 0-100. For example, if you perform prediction by training on a dataset that includes 3000 negative flares and 100 positive, the answer to this question would be 100/3100 = 0.0323. Percentage of Positive Samples = Positive Samples/Total Training Samples
  Quantifiable Response: Open.
- **Input Sample Size - Single Event Size:**
  Here you should answer with the information size (in bytes) for a single event. For example, if one positive or negative event is represented in your input dataset as a solar EUV image, then the answer to this question is the size of that image.
  Quantifiable Response: Open.
- **Time Coverage of Single Input Sample:**
  Here you should answer with the time (in hours) coverage of the single event you considered in the previous question. For example, if each one of your events is represented by a timeline, the answer to this question would be the amount of time this timeline covers. There might be works/models for which the answer to this question is zero, as they use for a single event only one solar imagery frame, or a single data point.
  Quantifiable Response: Open.

Output Information

Here we would like you to provide information about the testing/validation output your model provides. This section is related to the performance of your model and the type of predictions it offers.

- **Output Type - ML Prediction Category:**
  Here we would like to know what is the type of prediction your model performs. For example, some studies provide a binary prediction (Classification) whereas there are others that predict whether there are other studies that predict non-activity for the next X hours (All-Clear). Note here that a model can fall into multiple categories. For example, an LSTM/Regression model might predict a physical quantity's values (such as magnetic flux) for the next x hours, and then converts this predicted timeline to a probability of a positive event or an all-clear flag. In such a case, the Regression, Physical Quantity and Probability or All-Clear checkboxes must be selected. For more information on the output type categories please check Table 11 of Whitman et al. (2023).
  Categorical Response: Classification, Regression (Time-Series), Probability, Time Prediction (Onset Time/ Peak Time/ End Time), Physical Quantity Prediction (Peak/Fluence), All Clear, Other.

- **Triggered vs. Continuous Prediction:**
  This question is related to the previous question. Here we need to know whether the model outcome/prediction is Triggered or Continuous. A Triggered prediction means that something happens to the sun (flare, CME etc.) and there are parameters available that were not there before, and therefore prediction happens based on this event's parameters. An example would be a flare-based prediction. If your model relies on information from an event such as a flare or a CME, then it falls under the triggered category. A Continuous prediction model issues a warning at any time regardless of a solar precursor event as it relies on parameters available at all times. An example of a continuous prediction would be a regression model which provides continuous time-series prediction of a physical quantity, no matter whether there are flares or CMEs erupting.
  Categorical Response: Triggered, Continuous, Other.

- **Output Time Resolution - Forecast Window:**
  We define as forecast window the time period for which the forecast is valid, e.g. all clear for the next 24 hours. From another perspective, when a forecast is issued, the forecast window indicates the time period in which the predicted phenomenon is expected to occur, e.g. SEP threshold crossing in the next 7 hours.
  Quantifiable Response: Open.

Other Information

- **Relevant Publication:**
  Please add the DOI/Link to the publication this model is presented in (if published).
  Categorical Response: Open.

- **Developers:**
  Please add your name and any team members (if any) you would like me to include as a co-author.
  Categorical Response: Open.

- **SEP Event Definition:**
  Open ended. Please describe what is your SEP definition (ex. ≥ 10 MeV protons).
  Categorical Response: Open.

- **Validated using SEPVAL?**
  Let us know if you have used SEPVAL for the validation of your model or whether you would be interested in validating it for the review paper. The absence of common validation methods and therefore the difficulty of comparing the results of different models, will be discussed in the manuscript. The SEPVAL challenge time periods for 33 SEP events and 30 non-event periods can be downloaded from this Zenodo repository.
  Categorical Response: Yes, No, Other.

D. ACRONYMS

Acronym	**Meaning**
AA	Not defined (Model A.4)
AAS	American Astronomical Society
ACC	Accuracy
ACE	Advanced Composition Explorer
ADASYN	ADAptive SYNthetic
AR	Active Region
AUC	Area Under the Curve
AWT	Average Warning Time
BA	Balanced Accuracy
BBQ	Bayesian Binning Quantile
BiLSTM-SEP	Bidirectional LSTM - SEP (Model A.21)
BS	Brier Score
CANN	Custom Architecture Neural Network (Model A.13)
CART	Classification and Regression Tree (Model A.9)
CCMC	Community Coordinated Modeling Center
CDAW	Coordinated Data Analysis Workshop
CLEAR	Center for All-Clear Solar Energetic Particle Forecasts
CME	Coronal Mass Ejection
CNN	Convolutional Neural Networks
Cox PH	Cox Proportional Hazards
CSI	Critical Success Index
DOI	Digital Object Identifier
DONKI	Database Of Notifications, Knowledge, Information
DSCOVR	Deep Space Climate Observatory
ECE	Expected Calibration Error
EIT	Extreme-ultraviolet Imaging Telescope
EPI-Lo	Energetic Particle Instrument - Low
EPEAD	Energetic Proton, Electron and Alpha Detectors
EPREM-S	Energetic Particle Radiation Environment Module - S (Model A.24
EPS	Energetic Particles Sensors
ESA	European Space Agency
ESPERTA	Empirical model for Solar Proton Event Real Time Alert (Model A.5)
EUV	Extreme Ultra Violet
FAR	False Alarm Rate
FAR*	False Alarm Ratio
FM	Foundation Model
FPR	False Positive Rate
F1	F1 Score
GEO	Geostationary Earth Orbit
GeV	giga-electron Volt (unit)
GOES	Geostationary Operational Environmental Satellite
GridSearchCV	Grid Search Cross-Validation
GSS	Gilbert Skill Score
HEK	Heliophysics Event Knowledgebase

HMI	Helioseismic Magnetic Imager
HSS	Heidke Skill Score
IMAP	Interstellar Mapping and Acceleration Probe
IMP	Interplanetary Monitoring Platform
ISEP	Integrated Solar Energetic Proton Event Alert/Warning System
ISRO	Indian Space Research Organisation
IS⊙IS	Integrated Science Investigation of the Sun
I-ALiRT	IMAP Active Link for Real-Time
keV	kilo-electron Volt (unit)
KM	Kaplan–Meier
JSC	Johnson Space Center
LASCO	Large Angle and Spectrometric Coronagraph
LLM	Large Language Model
LOFAR	Low-Frequency Array
LSTM	Long Short-Term Memory
L1	Lagrange 1
L4	Lagrange 4
MAE	Mean Absolute Error
MCC	Matthew's Correlation Coefficient
MDI	Michelson Doppler Imager
MEMPSEP	Models for Probabilistic Forecast of Solar Energetic Particles (Model A.22)
ML	Machine Learning
MLP	Multi-Layer Perceptron
MLSW	(University of Michigan) Machine Learning for Space Weather[a]
MS-SEP	Not defined (Model A.8)
NASA	National Aeronautics and Space Administration
NCEI	National Centers for Environmental Information
NDA	Nançay Decameter Array
NN	Neural Network
NOAA	National Oceanic and Atmospheric Administration
PCC	Pearson Correlation Coefficient
pfu	particle flux unit ($counts/(cm^2 ssr)$)
PINN	Physics-Informed Neural Network
POD	Probability of Detection
PSP	Parker Solar Probe
PSPSP	PSP SEP Prediction (Model A.23)
PUNCH	Polarimeter to Unify the Corona and Heliosphere
REleASE	Relativistic Electron Alert System for Exploration
ReLU	Rectified Linear Unit
RH	Random Hivemind (Model A.10)
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
R^2	R^2 Score (Coefficient of Determination)
R2O	Research-to-Operations
SC	Solar Cycle
SDO	Solar Dynamics Observatory
SEM	Space Environment Monitor
SEP	Solar Energetic Particles
SEPEM	Solar Energetic Particle Environment Modeling
SEPVAL	Solar Energetic Particle Model Validation

[a] https://mlsw.engin.umich.edu/apps/runSEP

SEP-C	Not defined (Model A.12)
SEP-E	Not defined (Model A.14)
SHARP	Space Weather HMI Active Region Patches
SMARP	Space-Weather MDI Active Region Patches
SHINE	Solar Heliospheric and INterplanetary Environment
SILSO	Solar Influences Data Analysis Center
sklearn	Scikit Learn
SMOTE	Synthetic Minority Oversampling TEchniques
SOHO	Solar and Heliospheric Observatory
SolO	Solar Orbiter
SPAN-I	Solar Probe Analyzer for Ions
SPD	Solar Physics Division (refers to AAS)
SPE	Solar Proton Event
SPRINTS	Space Radiation Intelligence System (Model A.15)
SRAG	Space Radiation Analysis Group
SSEP	Survival SEP (Model A.11)
STEREO	Solar TErrestrial RElations Observatory
STSF	Supervised Time Series Forest (Model A.2)
SWEAP	Solar Wind Electrons Alphas and Protons
SWFO-L1	Space Weather Follow On – Lagrange 1
SWPC	Space Weather Prediction Center
S1	Minor Solar Radiation Storm as defined by NOAA[a]
S2	Moderate Solar Radiation Storm as defined by NOAA
S3	Strong Solar Radiation Storm as defined by NOAA
S4	Severe Solar Radiation Storm as defined by NOAA
S5	Extreme Solar Radiation Storm as defined by NOAA
TEBBS	Temperature and Emission measure-Based Background Subtraction
TSF	Time-Series Forest
TSS	True Skill Score
TS-HOG-TB	Time Series - Histogram of Oriented Gradients - TaBular (Model A.19)
UDM	Univariate Deep Merge (Model A.17)
UMASEP	University of MAlaga Solar particle Event Predictor (Model A.6)
UMASOD	University of Malaga predictor from Solar Data (Model A.7)
UNSPELL	UNifying Solar Particle Event modeLLing (Model A.18)
VAR	Vector Autoregression
XGBoost	eXtreme Gradient Boosting (Model A.1)
XRS	X-Ray Sensor
1D	1 Dimension
2D	2 Dimensions
3D	3 Dimensions

[a] https://www.swpc.noaa.gov/noaa-scales-explanation