



SEP Prediction Using the New SHMARP Dataset

Spiridon Kasapis¹, Irina Kitiashvili¹, Alexander Kosovichev², Viacheslav Sadykov³, Paul Kosovich², Patrick O'Keefe²

¹NASA Ames Research Center, ²New Jersey Institute of Technology, ³Georgia State University

Overview

- With human missions beyond Earth's magnetosphere planned, understanding and forecasting Solar Energetic Particle (SEP) events has become crucial.
- The SHMARP dataset merges data from Solar Cycles 23 and 24, utilizing both the SDO/HMI SHARP and SOHO/MDI SMARP datasets. The SHARP and SMARP are datasets that contain active region observations.
- Employed machine learning strategies, including Support Vector Machines (SVMs) and Regression Models, and achieved a prediction accuracy of 0.7 ± 0.1 , consistent with previous benchmarks¹, with a modest operational improvement noted.

SHMARP Dataset

- A new data product (SHMARP²) was created by filtering, rescaling and merging the SMARP and SHARP time series.
- The combined MDI-HMI dataset spans from April 1996 to May 2023, allowing for comprehensive temporal analysis and potential extension to current dates.
- Demonstrated significant correlations with the daily solar flare index and identified a leader-follower relationship, where certain parameters precede the flare index, indicating strong predictive potential.

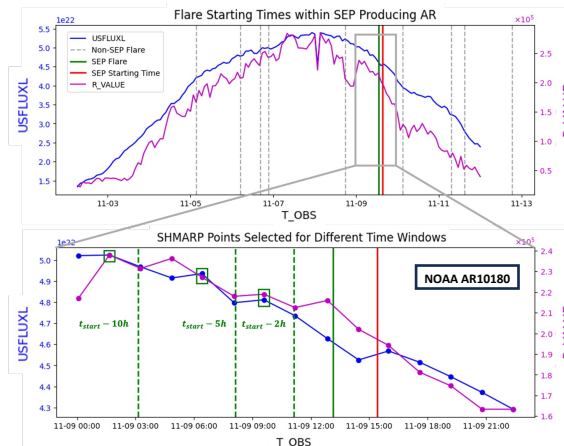
¹Kasapis, S. et al. (2022). Interpretable ML for SEP forecasts, Space Weather, 20, e2021SW002842.
²Kosovich, P. et al. (2023). Magnetic Field Parameters for Flare Forecasting. In Proc. of the 2023 Space Weather Workshop, Boulder, CO.
This research was supported by NASA Early Stage Innovation program grant 80NSSC20K0302, NASA LWS grant 80NSSC19K0068, NSF EarthCube grants 1639683, 1743321 and 1927578, and NSF grant 1835958.

Matching SEPs With Solar Flares

- We matched NOAA's database of 4238 flares with 168 SEP events recorded by NASA's Space Radiation Analysis Group, excluding A and B class flares.
- The SEP dataset was filtered to exclude unmatched SEPs which occurred on the far side of the solar disc or have missing information, resulting in a refined list of **110 Positive** and **3246 Negative** flares (3356 total).

SHMARP Data Selection and Split

Every single flare out of the 3356 flares has a corresponding SHMARP predictor vector, chosen from the entire SHMARP timelines as shown below.



Machine Learning Methods & Results

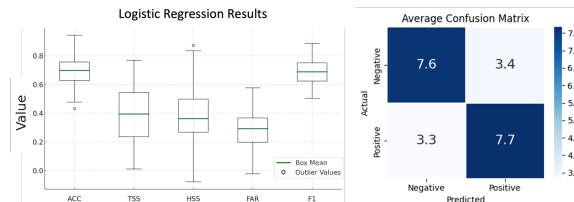
The SEP prediction capabilities of the SHMARP dataset is evaluated using SVM and Linear Regression models from Scikit-Learn, with a focus on exploring SVM kernels for data that may not be linearly separable and employing regression models suited for categorical (binary) outcomes, while hyperparameter tuning is conducted via GridSearchCV to optimize model performance.

SHMARP Predictors	SVM				Regression	
	Linear	RBF	Polynomial	Sigmoid	Logistic	Ridge
ALL PREDICTORS	0.64 ± 0.10	0.62 ± 0.10	0.64 ± 0.10	0.62 ± 0.10	0.65 ± 0.09	0.65 ± 0.12
USFLUXZ, R_VALUE, ANG_DIST_AR	0.66 ± 0.09	0.65 ± 0.10	0.66 ± 0.10	0.64 ± 0.10	0.65 ± 0.10	0.65 ± 0.09
USFLUXL, MEANGBZ, MEANGBZ	0.55 ± 0.10	0.55 ± 0.10	0.54 ± 0.10	0.53 ± 0.10	0.56 ± 0.10	0.56 ± 0.11
R_VALUE, ANG_DIST_AR	0.67 ± 0.10	0.66 ± 0.10	0.67 ± 0.11	0.67 ± 0.09	0.66 ± 0.10	0.65 ± 0.10
USFLUXZ, ANG_DIST_AR	0.66 ± 0.10	0.64 ± 0.11	0.64 ± 0.10	0.66 ± 0.10	0.66 ± 0.11	0.65 ± 0.10
USFLUXZ, R_VALUE	0.54 ± 0.11	0.58 ± 0.11	0.54 ± 0.08	0.58 ± 0.10	0.57 ± 0.09	0.55 ± 0.11

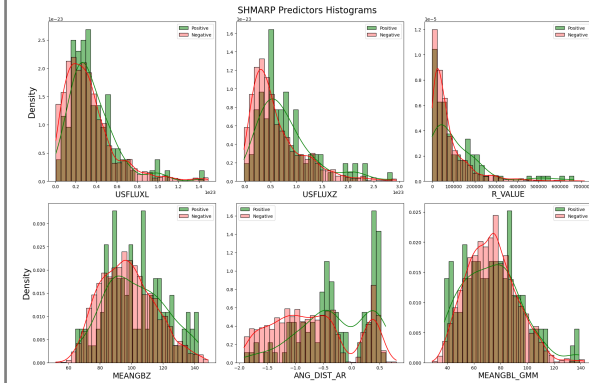
Accuracy (ACC) values for various models when trained on combinations of the six SHMARP predictors. For every value tabulated, 100 different runs with an equal amount of positive and negative data points (Balanced) are performed. Each model is performed using 90/10 train/test split.

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

Results for a variety of models when trained on the strongest predictors couple, the R_VALUE and the ANG_DIST_AR compared to the equivalent results by Kasapis et al. 2022. Below the detailed results of the best model (third row above).



SHMARP Predictors



Density histograms of the positive and negative SHMARP values for the six different predictors used in our ML models

Variable	Regular		Clean		Difference	
	t-statistic	p-value	t-statistic	p-value	t-statistic Δ	p-value Δ
USFLUXL	1.0418	0.2976	2.3254	0.0201	+1.2836	-0.2775
USFLUXZ	3.2500	0.0012	5.1861	2.31×10^{-7}	+1.9361	-0.0009
R_VALUE	2.9673	0.0030	6.4370	1.44×10^{-10}	+3.4697	-0.0029
MEANGBZ	2.9293	0.0034	4.2548	2.17×10^{-5}	+1.3255	-0.0032
ANG_DIST_AR	5.8973	4.07×10^{-9}	5.9810	2.51×10^{-9}	+0.0837	+0.0005
MEANGBL_GMM	-0.3880	0.6980	0.8734	0.3825	+1.2614	-0.3155

The t-statistic and p-value for the positive and negative SHMARP samples along with the histograms show the predictive capability of each physical quantity, agreeing with the machine learning results.

Conclusions

Using a 'Clean' dataset for training machine learning models results in slightly better SEP event prediction compared to using the 'Full' dataset, suggesting a potential advantage in focusing on event-specific data. Results remain consistent with previous research, indicating a stable accuracy ceiling despite larger data volumes. The work highlights the need for more complex analytical approaches and acknowledges the SHMARP dataset's role in enhancing understanding of SEP events and improving prediction models.