Solar Flare Forecasting using Machine Learning and SDO/HMI Data: A Comparison of Multiple ML Models and AR Parameter Time Series

Talwinder Singh[1], Christian Hall[2], Timothy Newman[2], Bernard Benson[3], Syed Raza[4], Nikolai Pogorelov[1,4]

[1]The Center for Space Plasma and Aeronomic Research, University of Alabama in Huntsville, Huntsville, AL, USA
[2]Department of Computer Science, University of Alabama in Huntsville, Huntsville, AL, USA
[3]McLeod Software Corporation, Birmingham, AL 35242, USA
[4]Department of Space Science, University of Alabama in Huntsville, Huntsville, AL, USA

Introduction

The primary objective of our work is to explore the potential of machine learning (ML) techniques in predicting solar flares, given their significant impacts on satellite communications, power grids, and Earth's climate. Accurate forecasting can help mitigate potential risks and improve preparedness for space weather events.

In this study, we have attempted to classify Active Regions (ARs) using a variety of ML classifiers, such as K-nearest neighbor (KNN), logistic regression (LR), random forest classifier (RFC), and support vector machine (SVM). We have investigated different aspects of solar flare prediction, including the optimal lead time before flare onset, the effect of excluding C class flares from the analysis, and the potential benefits of using time-series data of ARs.

Our initial findings suggest that predicting flare onset approximately 9 hours prior to the event yields promising results, especially with the RFC classifier. We also discovered that classifiers' performance improves when C class flares are excluded from the analysis. In addition, our research indicates that utilizing time-series data of ARs can significantly enhance solar flare prediction accuracy, with models achieving better performance when trained on time series of small subset of parameters rather than using point-in-time data of all parameters. Furthermore, we observed that the performance plateaus when training classifiers with up to 5 AR parameters, providing insights into the optimal combination of parameters for improved prediction.

Training Data

Database creation: We have created four lists of cases as shown here:

- A list of 767 M and X class flares.
- A list of 568 B and C class solar flares.
- A list of 1336 B class solar flares.
- A list of 622 ARs that did not flare when facing Earth.

All these lists had cases between years 2010 and 2021. We then created a database corresponding to the cases in these lists. The database contains magnetogram and intensitygram images for the 48-hour period before the onset of the flares in our lists. In the cases where ARs did not flare, we included in our database the data of these ARs 48 hours prior to them reaching the zero longitude in the Stonyhurst coordinate system. We have extracted HMI SHARP parameters from the fits files to create time series of these parameters for each of our cases. We have also calculated the flare predictive parameters defined by Korsos et al. (2016) for each of the magnetogram and intensitygram flares to use in our flare forecast models.

Data Correlation: We attempted to find correlation among the HMI SHARP parameters as part of an overall strategy to perform prediction in a reduced dimensional space. We looked at the data recorded at the start time of each B, C, M, and X class flare (similar studies were done for a 6h and 24h average before flare start time). A parameter was considered correlative if its the coefficient for two parameters was at least 0.7, in which case that parameter was dropped. Seven HMI SHARP parameters were concluded to be non-correlative: ABSNJZH, MEANGAM, MEANGAM, MEANJZD, MEANJZH, R_VALUE, and USFLUX.

Results using point in time data

We have conducted classifications of ARs into two categories using ML: those that produced flares above M class and those that generated flares weaker than M class or generated no flares. To accomplish this, we utilized a variety of ML classifiers, including KNN, LR, RFC, and SVM. Our classification approach initially involved analyzing point-in-time data prior to flare onset and examining the classification power at different lead times before the flare event. Figure 1 illustrates the performance of our models (true skill statistic) when trained on AR data obtained 24 hours (left panel) and 9 hours (right panel) before the flare onset. We observed the most favorable results with the RFC classifier. Moreover, we demonstrated that predicting subsequent flares from ARs that have already produced a solar flare earlier is more accurate compared to ARs that have not yet generated a solar flare.

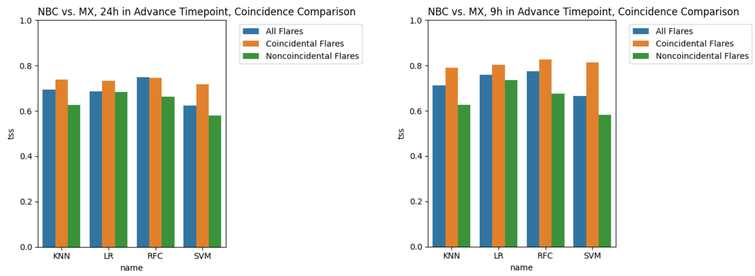

Figure 1: The performance of different classifiers (true skill statistic) using the AR data 24 hours (left panel) and 9 hours (right panel) before the flare onset. We also show that the performance is better for the ARs that have already flared within the previous 48 hours (coincident flares) as compared to ARs that did not flare in the last 48 hours (non-coincident flares)

We further investigated the optimal lead time before the flare onset by iterating our tests over a range of lead times. We found that the best performance is achieved with data approximately 9 hours prior to the event (see Figure 3 and the right panel of Figure 2).

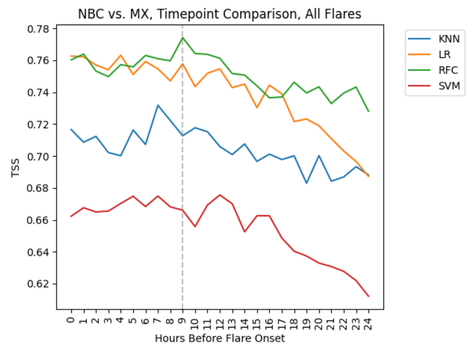

Figure 2: The performance of different classifiers (true skill statistic) using the AR data of all flares 0 to 24 hours before the flare onset. The best results are achieved by RFC when AR data 9 hours before the blare onset is used. This point is marked by the vertical dashed line in the figure.

The classifiers perform very well if we remove C class flares from our analysis, such that the classification becomes between MX class flare producing ARs and B class or no flare producing ARs. Following table shows the classification performance with different classifiers when using the database described in Sinha et al. (2022). This shows that the classifiers mainly struggle with the classification of ARs that produce C class flares.

Classifier	TSS
KNN	0.933
RFC	0.9492
LR	0.9339
SVM	0.9371

Results using time-series data

We have recently employed ML classifiers such as KNN, LR, RFC, and SVM to make predictions using time-series data of ARs. Our preliminary findings indicate that using time-series data for solar flare prediction yields promising results. In Figure 3, we compare the performance of our models when utilizing the full time series of the AR parameters between 13 to 24 hours prior to flare onset. We selected different triplets of AR parameters in this study and compared the classification performance in each case. Our analysis reveals that training the models with the time series of 3 parameters can yield better performance than using point-in-time data of all parameters.

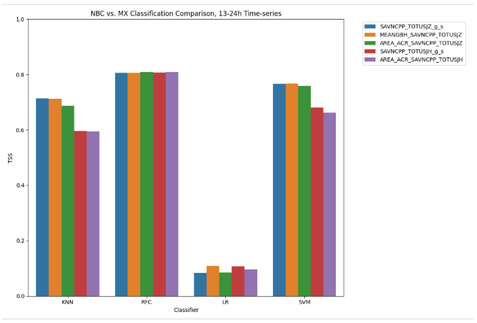

Figure 3: The performance of different classifiers (true skill statistic) when using time series of different triplets of AR parameters. RFC performs best across the board while LR performs very poorly when using the time-series.

We extended our analysis to include up to 5 AR parameters to train the classifiers. By testing multiple combinations of parameters, we concluded that the performance plateaus at 5 AR parameters. The following table shows how the performance (TSS) of the RFC classifier improves when increasing the number of SHARP parameters, it is trained on. These are the best results among all the possible parameter combinations.

SAVNCPP TOTUSJH TOTUSJZ	+d_I_f	+MEANJZH SHRGT45
0.818	0.831	0.835

Future work

We are currently expanding our time-series analysis by incorporating a diverse range of advanced ML algorithms. These include multi-layer optimal deep convolutional neural networks, Resnet-50, bidirectional long short-term memory networks, multi-layer perceptrons, gradient boosting decision trees, and dropouts meeting multiple additive regression trees. Our preliminary investigations with these models have yielded encouraging results, demonstrating the potential of these algorithms in enhancing solar flare prediction accuracy.

To further optimize our models, we plan to test them with various combinations of time-series windows, lead times, and parameter sets. We aim to identify the most effective combinations that can contribute to improved prediction performance.

Time-series analysis also introduces specific technical challenges, such as interpolation of missing data. To address this issue, we will conduct a comprehensive study on the effects of different interpolation techniques on model performance. By examining and comparing the impact of various interpolation methods, we hope to uncover insights that will help us refine our models and ensure more reliable predictions in the presence of missing or incomplete data.

References

- Korsos, M. B. and Erdelyi, R. (2016). On the State of a Solar Active Region Before Flares and CMEs., 823(2):153
- Sinha, S., Gupta, O., Singh, V., et al. 2022, ApJ, 935

Acknowledgement

This work was supported by NASA R2O2R GRANT 80NSSC22K0270