Investigating Performance Trends of Simulated Real-time Solar Flare Predictions: The Impacts of Training Windows, Data Volumes, and the Solar Cycle

Griffin T. Goodwin, Viacheslav M. Sadykov, Petrus C. Martens
Georgia State University

Contact: ggoodwin5@gsu.edu

Abstract

Key Questions: 1) How do standard machine learning classifiers used for flare prediction perform in real-time? **2)** How do training methodology, data volume, and solar activity affect forecasts?

Motivation: Flare prediction models are typically trained and tested using a random set of flaring (and non-flaring) data, which is inconsistent with real-time flare forecasting. Here, we focus on training classifiers using data only available prior to the forecast date.

Experiment: We train our classifiers with three different datasets selected from Georgia State's SWAN-SF database: **1)** a **stationary** window utilizing data prior to the *first* prediction in the series, **2)** a **rolling** window utilizing data from a constant time interval prior to the prediction, and **3)** an **expanding** window utilizing all data prior to the forecasting instance (see **Fig. 1**). We then investigate how performance scales with the number of features used, as well as the temporal size of the stationary and rolling windows (see **Fig. 2 & 3**). We also explore the relationship between classifier performance and the background soft X-ray (SXR) flux (see **Fig. 4**). Lastly, we develop an innovative method to visualize a classifier's performance as time progresses (see **Fig. 5**).

Salient Results: 1) Simple ML classifiers provide similar skill scores to more complex models when using point-in-time magnetogram data for real-time forecasts. **2)** In general, the number of features used does not have a significant effect on performance. **3)** When utilizing a 20-month stationary or rolling window, performance is comparable to the expanding window. A slight decrease in performance is observed when the window size is reduced. **4)** A strong positive Spearman correlation exists between the flare quiet false positive rate and the background SXR flux. High background flux complicates the detection of weak (~M1.0) flares and increases the potential for flares to overlap with stronger events in progress. Since our predictive models are based on magnetogram features, our models may be correctly predicting the occurrence of a flare. However, if the flare went undetected in SXR, it would be labeled incorrectly in SWAN-SF.

Methodology

Data

- Space Weather Analytics For Solar Flares (SWAN-SF)
  - Active region magnetogram time series data
    - Spans most of Solar Cycle 24 (2010 – 2018)
    - All time series are 12 hours in length with a 12-minute cadence.
    - 24 derived physics-based parameters
    - Labels are chosen based on the strongest flaring event in the following 24 hours.
  - Non-flaring events: Flare quiet + A, B, & C class flares
  - Flaring events: M & X class flares
  - To simplify our predictions, each time series is reduced to a single point-in-time summary statistic vector based on the mean, median, standard deviation, max, and min of each magnetogram parameter.
- Geostationary Operational Environmental Satellite (GOES) daily SXR flux data
  - We select the daily minimum value of the 1-8Å SXR flux as a proxy for the background SXR level.

Machine Learning Classifiers

- Decision Tree (DT)
- Support Vector Machine (SVM)
  - Gaussian radial basis function kernel
- Multilayer Perceptron (MLP)
  - Three-stage hidden layer (50 -> 25 -> 12 nodes)
- Hyper-parameters are optimized through an extensive grid search.
  - A stratified group 5-fold cross-validation is applied for parameter selection.

Methodology (Continued)

Simulated Real-time Training Windows

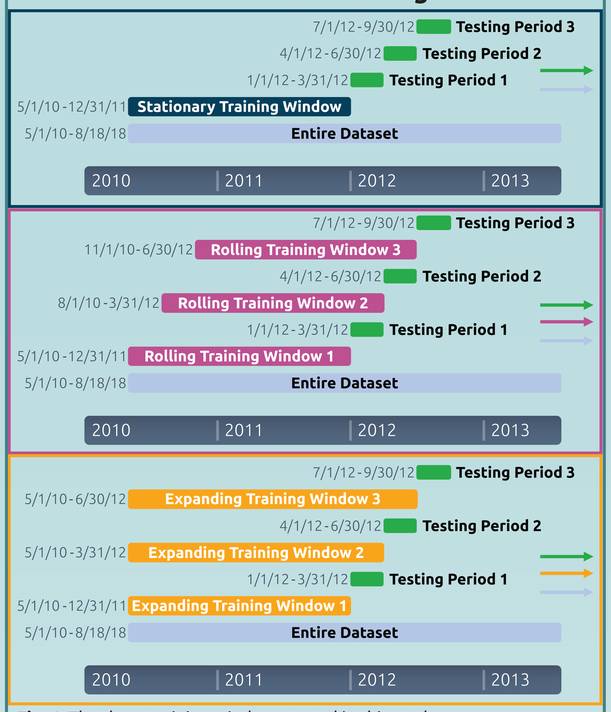

Fig. 1: The three training windows tested in this study.

- Training data is generated using one of the three windows shown in **Fig. 1**.
- Testing data is generated using three-month blocks following the first training window.
- We retain all flaring data and randomly undersample (while preserving climatology) non-flaring data within the training window to match the number of flaring events.
- To investigate how performance scales with the number of magnetogram features used, we test 1, 5, 10, 25, 50, and 120 features.
  - Features were selected based on those with the highest scoring ANOVA F-value in the training dataset.
- To investigate the impact of data volume on performance, we explore different stationary and rolling window sizes (5, 8, 11, 14, 17, and 20 months).

Performance Metrics

- True positives (TP), true negatives (TN), false positives (FP), false negatives (FN)

$$\text{True Skill Statistic (TSS)} = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$$

$$\text{Heidke Skill Score (HSS}_2\text{)} = \frac{2[(TP \times TN) - (FN \times FP)]}{(TP+FN)(FN+TN)+(TP+FP)(FP+TN)}$$

Background SXR Flux Correlation

- For each 25 feature classifier, we calculate the Spearman correlation coefficient between the monthly flare quiet false positive rate (FP / [FP + TN]) and the background SXR flux.

Training Window Results

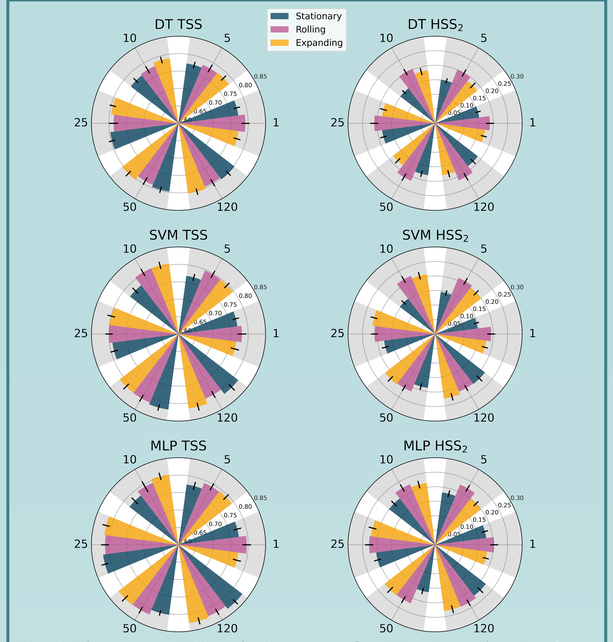

Fig. 2: The average TSS and HSS_2 scores for DT, SVM, and MLP using a number of features (1, 5, 10, 25, 50, 120) and window types. *Note: These results were obtained using a stationary and rolling window of 20 months.*

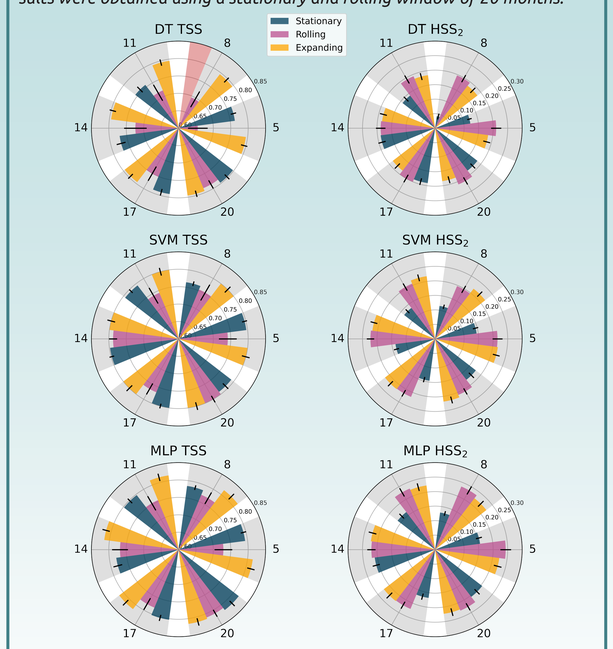

Fig. 3: The average TSS and HSS_2 scores for DT, SVM, and MLP (25 features) with varying window sizes (5, 8, 11, 14, 17, 20 months). Naturally, scores for the expanding window are the same across window sizes. *Note: The missing stationary data in the red wedge has an average TSS score of 0.26 +/- 0.05.*

SXR Correlation + Novel Visualization

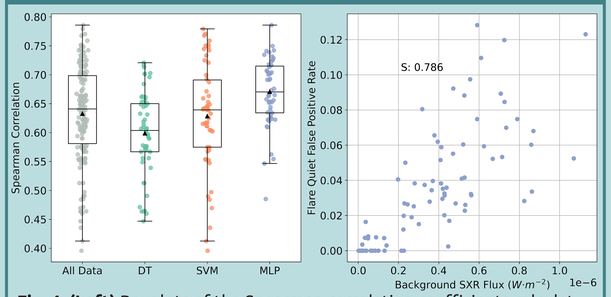

Fig. 4: (Left) Boxplots of the Spearman correlation coefficients calculated between the flare quiet false positive rate and the background SXR flux. The triangles indicate the mean of the distributions. **(Right)** A scatter plot for the strongest correlation observed in the MLP trials (14-month stationary window).

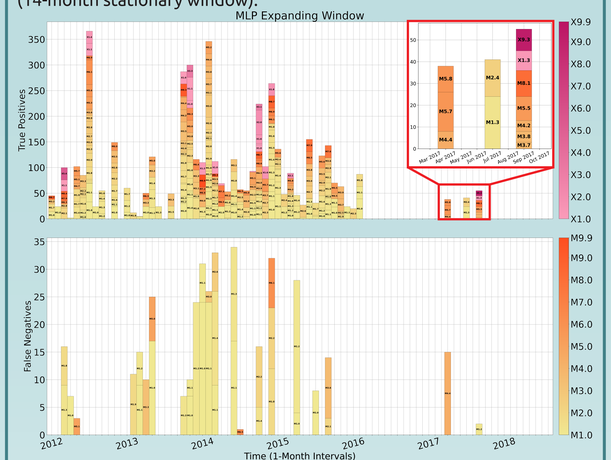

Fig. 5: A stacked bar chart of true positive and false negative predictions over time for a single trial of the MLP expanding window. Each bar represents flare counts grouped by flare strength, stacked over one-month intervals. Color corresponds to flare strength.

Conclusions

- 25 magnetogram features is optimal for balancing performance and complexity. Improvements beyond this threshold are minimal.
- Forecasting with a single feature does not significantly degrade classifier performance, emphasizing the inherent simplicity of our dataset.
- Interestingly, skill scores are similar across the 20-month stationary, rolling, and expanding windows. Below this threshold, TSS or HSS_2 scores gradually decrease, depending on the classifier and window type.
- DTs perform surprisingly well in comparison to SVMs and MLPs. This suggests that DTs are a viable alternative to these more complex models, especially if physically interpretable forecasts are important.
- To achieve the best possible performance, an MLP with a large rolling or expanding window and 25+ features should be implemented. If you are limited to a single training phase, a 20 month stationary MLP with 25+ features should be used.
- A strong positive Spearman correlation exists between the flare quiet false positive rate and the background SXR flux. This may be caused by obscured flares, which are incorrectly labeled as flare quiet in SWAN-SF.