Exploring Performance Trends of Simulated Real-time Solar Flare Predictions

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

Abstract

Key Question: How do standard machine learning classifiers used for flare prediction perform in real-time?

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. What happens if we train classifiers with data only available prior to the forecast date?

Experiment: We train our classifiers with three different data sets selected from Georgia State's SWAN-SF database: **1)** only data prior to the first prediction in the series, a "stationary" window, **2)** data from a constant time interval prior to the prediction, a "rolling" window, and **3)** all data prior to the forecasting instance, an "expanding" window (see **Figure 1**).

Visualization: We have developed an innovative method to visualize our forecasting performance that allows for the inspection of results, on an individual basis if needed, as time progresses (see **Figure 3**).

Salient Results: 1) To our surprise, skill scores only marginally improve for the expanding training window. In other words, training with more flares does not significantly increase the quality of predictions (see **Table 1/2 and Figure 2**). **2)** Through our visualization and Pearson correlation statistics, we determined that the background soft X-ray flux and the solar cycle strongly influence the performance of the classifiers. High background flux complicates the detection of weak (~M1.0) flares and increases the potential for flares to overlap with stronger events in progress. We observe this in **Figure 3 / Table 4** with the percent of flare quiet events forecasted as flaring increasing during periods of high background flux. This suggests that these flare quiet regions could be mislabeled in the SWAN-SF data set. Additionally, we find that classifier performance tends to decrease when approaching solar maximum.

This research is supported by NASA LWS Grant 80NSSC22K0272.

Introduction

What Are Solar Flares?

- Bursts of electromagnetic radiation
- Originate from magnetic active regions
- Logarithmic measurement scale
  - A (weakest), B, C, M, X (strongest)
- Pose a massive threat to astronauts and electronics

How Are They Studied?

- Machine learning models
- Individual predictions (true positives - **TP**, true negatives - **TN**, false positives - **FP**, false negatives - **FN**)
- True skill statistic (TSS)
  - $TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$
- Heidke skill score (HSS2)
  - $HSS2 = \frac{2*[(TP*TN)-(FN*FP)]}{(TP+FN)*(FN+TN)+(TP+FP)*(FP+TN)}$

Goals of This Study

- Investigate the performance of simulated real-time predictions
- Determine any prediction dependencies
  - Training window type
  - Solar background X-ray flux
  - Solar cycle
- Improve visualizations

Methodology

Data

- Space Weather Analytics For Solar Flares (SWAN-SF)
  - Spans part of solar cycle 24 (2010 – 2018)
  - Magnetogram time series data of active regions
    - Each 12 hours in length
    - Labeled based on the strongest flaring event in the following 24 hours
  - Quiet, A, B, C – Labeled as non-flaring event
  - M, X – Labeled as flaring event
  - To simplify our prediction, the time series data was reduced to a single multidimensional point based on the time series summary statistics
- Geostationary Operational Environmental Satellite (GOES) daily X-ray flux data
  - Utilized 1-8Å SXR channel
  - Selected minimum flux for each day as a proxy for background level
- Sunspot Index and Long-term Solar Observations (SILSO) daily sunspot data

Machine Learning Classifiers

- Decision Tree (DT)
- Support Vector Machine (SVM)
  - Gaussian radial basis function kernel
- Feed-Forward Neural Network (NN)
  - Three hidden layers (100 → 50 → 25 nodes)
  - 50% dropout rate
- Optimized through grid search / trial and error
- All classifiers were trained using the 20 magnetogram features with the highest ANOVA F-values

Simulated Real-time Training Windows

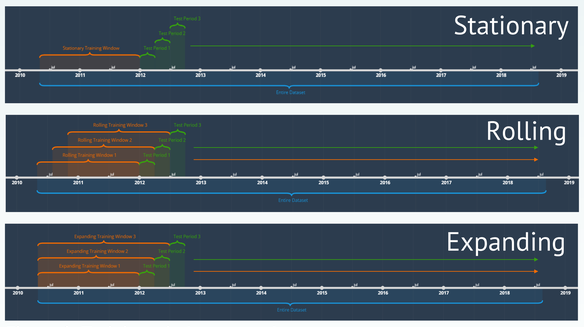


Figure 1: Training windows tested

Background X-ray Flux Dependency

- Calculate the Pearson correlation coefficient between the background X-ray flux and the percent of quiet events labeled as flaring based on the expectation that flares are obscured during periods of high background X-ray flux.

Solar Cycle Dependency

- Calculate the Pearson correlation coefficient between the TSS scores and the average number of sunspots during testing.

Training Window Results

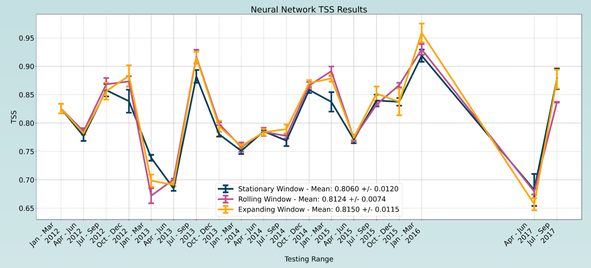


Figure 2: Neural network TSS vs. testing period

	Stationary	Rolling	Expanding
DT	0.7974 ± 0.0195	0.7817 ± 0.0309	0.7920 ± 0.0180
SVM	0.7984 ± 0.0201	0.8049 ± 0.0054	0.8124 ± 0.0115
NN	0.8060 ± 0.0120	0.8124 ± 0.0074	0.8150 ± 0.0115

Table 1: Average TSS scores for different classifier and window combinations across three trials

	Stationary	Rolling	Expanding
DT	0.1742 ± 0.0091	0.1983 ± 0.0156	0.1715 ± 0.0305
SVM	0.1504 ± 0.0071	0.1993 ± 0.0111	0.1991 ± 0.0141
NN	0.1746 ± 0.0159	0.1974 ± 0.0116	0.2097 ± 0.0236

Table 2: Average HSS2 scores for different classifier and window combinations across three trials

	Stationary	Rolling	Expanding
DT	37,097.67 ± 131.28	35,101.33 ± 135.31	39,831.66 ± 199.42
SVM	45,375.67 ± 60.85	31,830.00 ± 82.60	33,350.67 ± 151.64
NN	39,381.33 ± 137.15	32,722.67 ± 91.88	32,074.00 ± 150.57

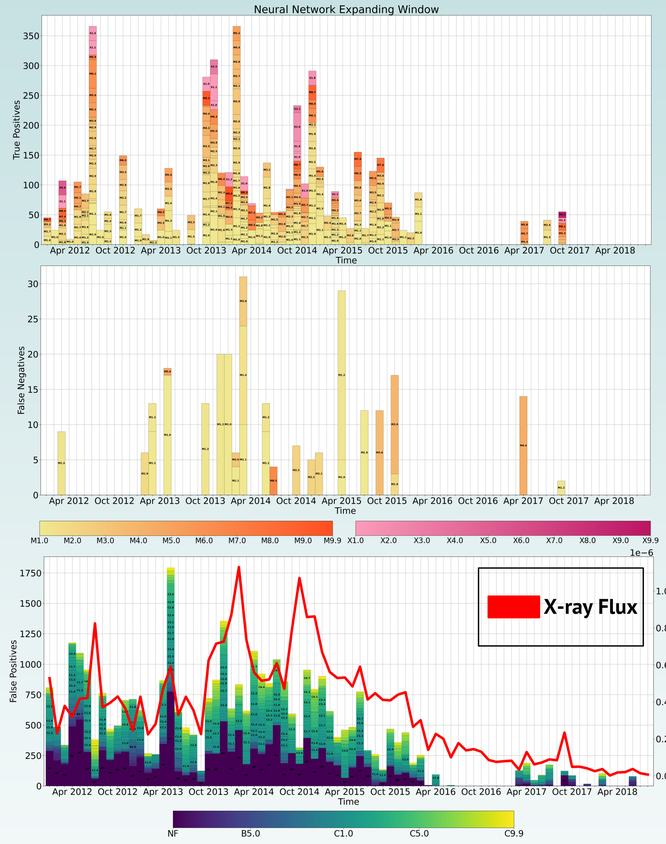
	Stationary	Rolling	Expanding
DT	253.67 ± 4.71	307.00 ± 5.38	193.67 ± 5.28
SVM	99.67 ± 2.10	293.00 ± 1.27	248.00 ± 3.73
NN	171.00 ± 2.70	255.00 ± 1.86	268.00 ± 2.62

Table 3: Average number of false positives (top) and false negatives (bottom) across three trials

NOTE: There are a total of 267,108 test data points. 5,141 flaring and 261,967 non-flaring

New Visualization, X-ray Flux, and The Solar Cycle

Figure 3: A stacked bar chart of TP, FN, & FP counts for the neural network expanding window versus time. Each bar represents flare counts based on strength, stacked over one-month intervals. Color corresponds to flare strength.



	Stationary	Rolling	Expanding
DT	0.6476	0.5754	0.4896
SVM	0.6093	0.4810	0.4830
NN	0.6205	0.6025	0.5855

Stationary	Rolling	Expanding
-0.5925	-0.6181	-0.5701
-0.5939	-0.4163	-0.5445
-0.5594	-0.4891	-0.5171

Table 4: Pearson correlation coefficient between background X-ray flux and the percent of flare quiet false positives (left) / TSS scores and average number of sunspots during testing (right)

Conclusions

- The performance of the classifiers tested does not differ significantly across all three training windows.
- Surprisingly, the stationary training window performs best for "all-clear" predictions.
- The soft X-ray background flux strongly affects forecasting quality, with an increase in the percent of supposed flare quiet regions predicted as flaring during periods of high background flux. This suggests that some of these regions may be mislabeled in the SWAN-SF data set.
- Classifiers tend to perform worse as they approach solar maximum.