Understanding the Impact of Statistical Time Series Features for Flare Prediction Analysis

Maxwell Hostetter*, Azim Ahmadzadeh*, Berkay Aydin*,
Manolis K. Georgoulis[†‡], Dustin J. Kempton*, and Rafal A. Angryk*
*Department of Compute Science, Georgia State University, Atlanta, GA, USA
[†]Department of Physics & Astronomy, Georgia State University, Atlanta, GA, USA
[‡]RCAAM of the Academy of Athens, Athens, Greece
Email: *{mhostetter1,aahmadzadeh1,baydin2,dkempton1,angryk}@cs.gsu.edu,
[†]manolis.georgoulis@phy-astr.gsu.edu

***Abstract*—Machine learning-based space weather analytics has attracted much attention due to the potential damages that can be caused by the extreme space weather events. Using a recently released data benchmark, named SWAN-SF, designed for solar flare forecasting based on the pre-flare time series of solar magnetic field parameters, we conduct a case study on the impacts of statistical features derived from the multivariate time series. We investigate the relationship between the number of needed statistical features extracted from the multi-variate time series and the performance of flare forecast models. To that end, we employ random forest and mean decrease impurity to determine a feature selection methodology along with an evaluation procedure. The proposed evaluation method delivers a balance between the two frequently used metrics in this domain, namely True Skill Statistic and Heidke Skill Score. Our approach allows to introduce a generic feature selection and evaluation procedure that is independent from the minor and often obscured decisions that must be made for having a binary forecast model, while presenting interpretable and actionable tools that can help non-data experts make more informed and realistic decisions.**

***Index Terms*—time series classification, feature selection, statistical time series features, flare prediction**

I. Introduction

Our society is characterized by a complex interweave of interdependencies among its critical infrastructures. Severe space weather events such as large solar flares and coronal mass ejections, can have tremendous impact on our increasingly technologically-dependent society [1]. The socioeconomic impact of space weather events includes not only the industry-specific events (such as spacecraft anomalies, power outages, or aircraft re-routing) but also the collateral effects of technology failures on dependent infrastructures and services. Many studies from governmental and independent institutions confirm the existence of these impacts and estimates the accrued economic damages [2]–[4].

Extreme space weather events are low-frequency but high-consequence events [5] and therefore, in terms of their potential broader, collateral impacts, present a unique set of problems for public and private institutions and governance, different from the problems raised by conventional, expected, and frequently experienced events. A workshop report from National Research Council [1] suggests that a contemporary repetition of the 1859 Carrington Event [6] would cause extensive social and economic disruptions. This strongly motivates the need to understand the mechanisms behind extreme space weather events, i.e. flares, coronal mass ejections, and solar energetic particle events, and anticipate/forecast their occurrences in advance to reduce the potential socioeconomic upheaval and damage caused.

Our aim is to facilitate the multivariate time series classification for solar flare prediction by discovering the important precursors for these low-frequency and high-risk events. We present a feature selection method utilizing built-in Random Forest (RF) ranking, which is specifically tuned for time series-based solar flare prediction algorithms. Multivariate time series classification can be performed with recurrent neural networks [7] or their contemporary counterparts such as long short-term memory networks [8]; however, our focus is creating a physical understanding, which requires us to explore and carefully analyze the evolution of time series rather than attempting to beat other flare prediction techniques. To understand the evolution, we will extract a large number of statistical features from each univariate time series that constitute multivariate time series instances. In a nutshell, our method can be seen as an embedded, high dimensional feature selection technique for rare event prediction.

II. Related Work

Schrijver et al. quantified the association of flares with the total unsigned flux around high gradient, strong-field polarity-inversion lines (PILs). This parameter is called R–value, and it was observed that no X- or M-class flares occur within 24 hours of the determination of R–value for all the cases where R–value < 2.8 [9]. From the opposite angle, it was observed that only 9% of the active regions with R–value > 2.8 experienced X- or M-class flares with the same observation window. These findings result in the R–value to become a convenient classification tool, that can be used for forecast of strong flares.

In another attempt to forecast flares, Bobra et al. [10] used a machine learning classifier on four years of data gathered by the Solar Dynamics Observatory's (SDO) Helioseismic and Magnetic Imager (HMI) [11]. They achieved a relatively high forecast performance in their work. What differentiates their

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

results from others', however, is not the model's performance. It is first, the size of the data used, i.e., four years worth of Spaceweather HMI Active Region Patch (SHARP) vector magnetic field data (2010 through 2014), and second, a list of 25 physical parameters employed, derived from active regions, calculated every 12 minutes during the lifetime of each active region. R–value, already proven to be a powerful predictor for flares [9], is only one of these 25 parameters, which emphasizes the robustness of the forecast and reliability of the findings.

More recently on this topic, Campi et al. [12] collected a much larger list of parameters. Utilizing nearly 200 features extracted from a variety of physical parameters, developed within the Horizon 2020 FLARECAST project[1], this work introduced a comprehensive analysis on this topic with the highest dimensionality ever used. To mitigate the curse of dimensionality, a recursive feature elimination methodology was employed to also take into account the correlation between the features in the elimination process. In this study, in order to determine the features ranking, they used two models: hybrid LASSO (HLA) and RF, and compared their performance by thresholding on the probabilistic outcome of the classifiers. Although, their final results in terms of True Skill Statistic (TSS) and Heidke Skill Score (HSS), are systematically lower than those achieved and reported by [10] and [13], the performance seems to be closer to reality. This is evident in their sampling methodology and the fact that they avoided presence of the identical feature vectors in both training and testing sets.

The RF classifier, next to Support Vector Machine (SVM), has become a very popular tool in flare forecast studies, in particular in ranking of the features in terms of their contribution to the forecast. Florios et al. [13] also used this classifier. They worked toward forecast of solar flares taking both probabilistic and dichotomous approaches into account. The experiments were carried out on more than 23,000 observations of 7 parameters computed using either the line-of-sight magnetograms of SHARP data [14] or respective radial components, as in [10]. They experimented with Multi-layer Perceptrons (MLP), SVM, and RF as their binary classifiers, and Linear Regression, Probit Regression, and Logistic Regression, as the probabilistic models, with different probability thresholds. They concluded that RF performs better than all the others, in terms of TSS and HSS. Of course, with different configurations and datasets that are computed based on different features, or collected differently, the results may vary. However, a benefit of RF lies in its built-in feature ranking mechanism. This allows a multivariate feature selection tool whose results could be interpreted and therefore be valuable for solar physicists. This is in particular important because instead of a black-box forecast model, the RF may provide insight to the problem.

Another flare forecast study conducted by Domijan et al. [15], on the features extracted from line-of-sight magnetograms (from SOHO/MDI) by Solar Monitor Active Region Tracker (SMART) algorithm [16], also employed RF to help rank their features. In addition, a Maximal Marginal Relevance (MMR) filter, and Lasso [17] were utilized to handle a multivariate feature selection, and a set of classifiers, namely Logistic Regression, Support Vector Machines, and Deep Neural Networks, were used to guide the feature selection process by iterative evaluation of the classification task.

The studies reviewed above are just a few of many valuable scientific investigations on solar flare forecasting. What differentiates our analytic is the dataset we run our experiments on. As we discuss in the following section, to the best of our knowledge, this is the first data benchmark that focuses on time series of pre-flare characteristics of magnetic field within active regions, rather than point-in-time data points. Using RF for investigation in the importance of the physical parameters allows a probabilistic approach which provides more flexibility and interpretability to the findings. We try to establish a sound methodology for reducing the number of features and training a flare forecast model toward the end of determining important physical parameters whose pre-flare behavior could help indicate a solar flare. We discuss the challenges and the results in Sec.V.

III. Dataset

In this section, we discuss the data benchmark used to build our dataset suitable for the specific objectives outlines before. We briefly talk about the original data, and its collection process, and then we elaborate on the generated dataset that we use for our experiments.

A. *SWAN-SF Data Benchmark*

In this study, we use a recently created data benchmark, named as Space Weather ANalytics for Solar Flares (SWAN-SF), made entirely of multivariate time series, aiming to carry out an unbiased flare forecasting. We hope that rigorous analyses on this data benchmark could open new doors for flare forecast studies, and set at rest at least some of the important questions that solar physicists have been investigating for the past century.

SWAN-SF comprises five partitions which are temporally non-overlapping. The partitioning is carried out in such a way that each of them contains approximately an equal number of X- and M-class flares. Distribution of the classes in each partition can be seen in 1. The data points are time series slices of 24 physical (magnetic field) parameters extracted from the flaring and non-flaring regions, in a sliding fashion. That is, for a particular reported flare, corresponding to an active region with a unique id, k equal-length multivariate time series are collected using a temporal window sliding over the history of the flare, extracting physical parameters from its active region. This is called an *observation window* denoted by T_{obs}. It spans over 24 hours of flares' history. Given that t_i indicates the starting point of the i-th slice of the multivariate time series, the $(i+1)$-th slice starts at $t_i + \tau$, where $T_{obs} = 8\tau$.

The four different flare classes considered in this benchmark are X, M, C, and B, and with the addition of the time series

[1] http://flarecast.eu

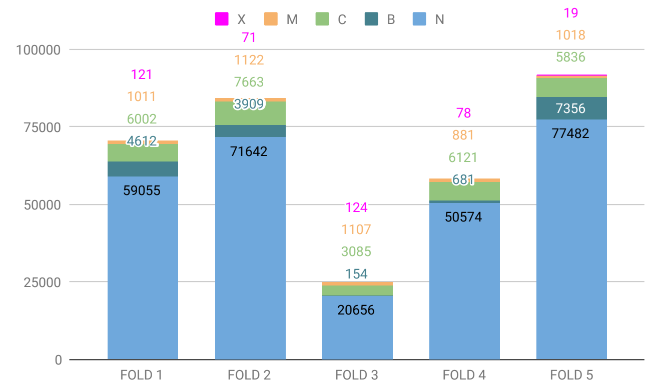

Fig. 1. Distribution of flares of different classes in the SWAN-SF dataset.

extracted from the flare-quiet regions (labeled as N), form a 5-class dataset. In this study, we simplify the task by merging the stronger instances (i.e., X- and M-class flares) to form the positive class, and the weaker instances (i.e., of C, B, and N classes) to represent the negative class. Throughout this text, we refer to the positive and negative classes as *flaring* and *non-flaring*, respectively.

Generally in machine learning, there are two main approaches in working with the time series data. One is to preprocess the time series and feed them directly into the statistical models. The other is to extract a set of statistical features from the time series and use their descriptors instead of the actual time series for the models to learn from. Our primary goal of finding the most effective features in prediction of flares, guides us to use the former, i.e., the extracted features. Following, we elaborate more on the statistical features employed in this study.

B. Statistical Features

To the best of our knowledge, SWAN-SF data is the first data benchmark that focuses on time series, rather than point-in-time data points for a period just short of a decade. Therefore, perhaps except in a few instances, there are no established theories as to which characteristics of the time series may show a significant flare-predictive capability. This encourages us to hand-pick, from a plethora of trend analysis studies, a set of time series features as the descriptors that could potentially add to the forecasting power of the predictive models. Working on the extracted features of time series also has the advantage of reducing the dimensionality of the data, from that being the length of the time series, to the number of features.

The chosen statistical features are listed in Table I. In general, these 43 features can be grouped into several clusters of descriptors. In the first and second clusters, there are the descriptive statistics (e.g., *min*, *max*, etc) that describe the time series in their entirety and also those that compare the first and second halves of the time series. The third group contains the representation of the time series in terms of their extrema (e.g., *number of local minima*, *average of local minima*). We also include several derivative-based features using the difference derivative and gradient derivatives, and form the forth and fifth groups, respectively. Another collection, i.e., the sixth one, is wrapped around the features that quantify the high-level changes within the time series (e.g., *linear* and *quadratic weighted average*s, or *average absolute change*). The seventh group describes the positive and negative fractions of the time series, which disregards the temporal aspect of the time series, whereas the features in the eighth group, that focus on the tail of the time series (e.g., *last value*). And finally, the ninth cluster contains several features describing the long-run trends of the time series (e.g., *slope of longest monotonic increase)*.

The final dataset has a dimensionality of 1032, with 43 statistical features extracted from 24 physical parameters. Data points of this dataset are labeled by 5 different classes of flares, namely GOES X, M, C, B, and N. The latter represents flare-quite instances and also contains GOES A-class events.

IV. Methodology

A. Data Preprocessing

After accessing the data, we calculate the statistical features of the time series to get our transformed data set. After computing the statistical features, the dataset requires a minimal preprocessing due to the presence of some missing values. Since this accounts for a very small fraction of the data (i.e., $< 0.01\%$), we simply utilize linear interpolation to reproduce those values.

As mentioned in Section 3, we merge the labels N, B, and C into the negative class and M, and X into the positive class.

B. Model Selection

RF [18] is an ensemble model, made by aggregating a number of decision trees. Figure 2 shows an example decision tree of depth 1. In the top of the tree is the root node where a splitting criterion is specified. The impurity of the node is measured by the gini index [19], the number of samples in each node is reported, as well as the number of samples from each class. The nodes at the bottom of the tree are leaf nodes, a label is assigned based on the class with majority weight in a node. During construction, each tree is trained on a bootstrap sample of the training set. We can constrain the number of features considered at each node when looking for the best split. During prediction, each tree is applied to a new observation. The observation ends up in a leaf node of the tree, which contains the proportion of each class that reach that leaf node during training. These proportions are issued as a pair of class probabilities. Since these probabilities sum to 1, we can consider only the probability that an observation is in the positive class. These probabilities are averaged over all trees to give the predicted probability of a sample being in the positive, in our case, flaring class.

We chose RF for a number of reasons. The building blocks of RF–decision trees–are highly interpretable, and we want our results to provide insights to domain experts. We are able to use original feature values rather than normalizing the data

TABLE I
LIST OF ALL STATISTICAL FEATURES EXTRACTED FROM THE TIME SERIES

Group	Features	Description
1	$min(ts)$, $max(ts)$, $median(ts)$, $\mu(ts)$, $\sigma(ts)$, $skewness(ts)$, $kurtosis(ts)$	descriptive statistics on the time series ts
2	$min(ts^{\dashv}) - min(ts^{\vdash})$, $max(ts^{\dashv}) - max(ts^{\vdash})$, $med(ts^{\dashv}) - med(ts^{\vdash})$, $\sigma(ts^{\dashv}) - \sigma(ts^{\vdash})$, $sk(ts^{\dashv}) - sk(ts^{\vdash})$, $ku(ts^{\dashv}) - ku(ts^{\vdash})$	differences between the descriptive statistics on the first and the second half of the time series ts
3	$\|\{local\_minima\}\|$, $\|\{local\_maxima\}\|$, $\|\{local\_extrema\}\|$, $\|\{zero\_crossings\}\|$, $\mu(\{local\_mimima\})$, $\mu(\{local\_maxima\})$, $\mu(\{local\_maxima\_upsurges\})$, $\mu(\{local\_minima\_downslides\})$	representation of time series in form of their extrema
4	$\mu(ts')$, $\sigma(ts')$, $skewness(ts')$, $kurtosis(ts')$	descriptive statistics on derivative (i.e., windowing differences) of the time series
5	$\mu(\partial ts), \sigma(\partial ts), \sigma^2(\partial ts)$, $skewnesss(\partial ts)$, $kurtosis(\partial ts)$	descriptive statistics on derivative (i.e., approximation of analytic gradient) of the time series
6	$lwa(ts)$, $qwa(ts)$, $\mu(abs(ts'))$, $\mu(abs(\partial ts))$	linear and quadratic weighted average of times series, and changes of the derivatives
7	$\frac{\|\{p \in ts; p>0\}\|}{n}$, $\frac{\|\{p \in ts; p<0\}\|}{n}$	positive and negative fractions of records in a times series of length n
8	$lv_{k=1}(ts)$, $\sum(lv_{k=10}(ts))$, $\mu(lv_{k=10}(ts))$	description of time series in terms of their k last values $(lv_k(ts))$
9	$longest\_positive\_run$, $longest\_negative\_run$, $longest\_monotonic\_increase$, $longest\_monotonic\_decrease$, $slope(longest\_monotonic\_increase)$, $slope(longest\_monotonic\_decrease)$, $\mu(\{slope(monotonic\_increases)\})$, $\mu(\{slope(monotonic\_decreases)\})$	long-run trends of the time series ts

Notations. μ: mean, σ: standard deviation, sk: skewness, ku: kurtosis, ts: time series, $ts^{\dashv}$: first half of ts, $ts^{\vdash}$: second half of ts, $||$: set cardinality, ts': difference derivative of ts, ∂ts: gradient derivative of ts, lwa: linear weighted average, qwa: quadratic weighted average, $abs(ts)$: absolute value of each ts_i, $lv_k(ts)$: last k values of ts.

which would present a new set of challenges as discussed in [20]. As well, RF gives us access to a built-in feature ranking method, which we can compare to univariate feature ranking.

C. Hyper-parameter Tuning

We perform hyper-parameter tuning to determine suitable settings for the model before running experiments. During hyper-parameter tuning, we tune the maximum depth of the trees, class-imbalance solutions, and the number of features randomly selected at each node. We perform cross validation by forming a training set of four partitions and leave one partition out as the test set. We measure performance with a metric called *average precision* which is detailed in Section IV-E. We consider the set of parameters which yielded the highest mean performance when tested across all partitions to perform best.

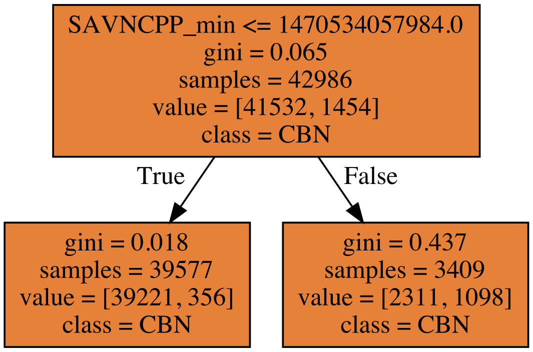

Fig. 2. An example of a Decision Tree of depth 1, with the gini scores and population sizes both before and after a split. The parameter used in the root node is *SAVNCPP* that stands for *Sum of the Absolute Value of the Net Current Per Polarity*.

D. Feature Selection

We perform feature selection to increase the interpretability of our model and improve predictive performance.

For feature selection, we consider one method given by RF and a univariate selection method. Univariate feature selection is done a priori on the entire dataset and yields a single ranking of all features. When using RF based method, a model must first be trained. As in hyper-parameter tuning, we form a training set of four partitions and leave one partition out. After training, we compute a feature ranking. Features are ranked by the standard deviation of the mean decrease impurity (MDI) which is detailed in IV-E; a feature with a smaller standard deviation of the MDI will be ranked higher than a feature with a larger standard deviation of MDI. We then repeat the process, this time leaving out a different partition. We compute five feature rankings with this way. We then aggregate the results by averaging over standard deviation of MDI.

E. Metrics

To rank features or to compare performance of different models, we need to first define a few metrics.

For one method of feature ranking, we use the Fischer score (F-score). The F-score is a univariate feature score that does not consider interdependencies among features, but does give a score for each feature based on how well they can separate the two classes. The formula for F-score is as follows:

$$F(i) = \frac{(\bar{x}_i^+ - \bar{x}_i)^2 + (\bar{x}_i^- - \bar{x}_i)^2}{\frac{1}{n^+ - 1}\sum_{k=1}^{n^+}(\bar{x}_{k,i}^+ - \bar{x}_i)^2 + \frac{1}{n^- - 1}\sum_{k=1}^{n^-}(\bar{x}_{k,i}^- - \bar{x}_i)^2}$$

where n^+, n^- indicate members of the positive and negative class respectively, $\bar{x}_i$ is the mean value of feature i, and $\bar{x}_i^+$

and $\bar{x}_i^-$ are the mean over the positive and negative samples of i-th feature respectively.

From RF, we get access to a measure of variable importance based impurity: a group of samples at a node in a decision tree with an even number of members from each class would be considered impure, samples with members from only one class be considered pure. We measure impurity with the gini index,

$$Gini = 1 - \sum_{c \in Classes} p_c^2$$

where p_c is the probability of randomly selecting a sample from class c.

Each time a split is formed in a decision tree, the impurity from one node to the next should decrease. We can rank features based on how well they can be used to form splits. We use Mean Decrease Impurity (MDI) for this. MDI is detailed in [21],

$$Imp(X_j) = \frac{1}{M} \sum_{m=1}^{M} \sum_{t \in \varphi_m} \mathbf{1}(j_t = j)[p(t)\Delta i(s_t, t)]$$

where M is the number of estimators, j_t is the feature used for splitting at node t, $p(t)$ is the proportion of samples reaching node t, $\Delta i(s_t, t)$ is the change in impurity caused by the split s_t at node t, and $\mathbf{1}(j_t = j)$ is the indicator function.

When we perform binary classification, each predicted result will fall into one of four categories: A flare that is classified as a flare is considered a True Positive (TP), a flare misclassified as non-flaring is a False Negative (FN), a non-flare classified as non-flare is a True Negative (TN), and a non-flare classified as a flare is a False Positive (FP).

Two common performance metrics used in flare prediction for binary classification are the previously mentioned TSS and HSS. Use of these metrics is recommended in [22].

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN}$$

$$HSS = \frac{2 \times [(TP \times TN) - (FN \times FP)]}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}$$

TSS has the benefit of being unbiased with respect to class imbalance and is useful for comparison between experiments. It has a range of -1 to 1. A forecast with no mistakes receives a 1 and forecasts that are random or constant score a 0. HSS measures the improvement of the model over random chance, but is affected by changes in the class imbalance ratio. It has a range of $[-\infty, 1]$, where negative values indicate the forecast is worse than chance, 0 is random or no skill, and 1 is a perfect forecast.

As mentioned in Section IV-B, the output of our RF will be a probability. To convert this to a binary classification, we must set a probability threshold, over which a sample will be considered as flaring.

Because our dataset has a severe class-imbalance issue, precision and recall are good candidates for measures of binary classification performance. Neither measure includes TN which could potentially be much larger than than the sum of TP, FP, and FN and cause a metric to give a misleading result.

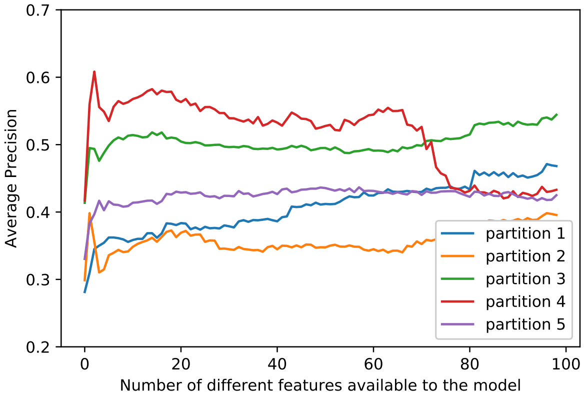

Fig. 3. Model performance in terms of Average Precision with varying number of features selected based on MDI.

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

By varying the probability threshold in the range $[0, 1)$, we generate a series of precision and recall pairs. These pairs are used to plot the precision-recall curve. The precision-recall curve then, is a way to examine model performance across all probability thresholds. We use average precision (AP) to summarize this curve,

$$AP = \sum_i (R_i - R_{i-1}) P_i$$

where R_i, P_i are the recall and precision resulting from the i-th threshold.

We use average precision as the criterion for ranking model performance. We do this because RF outputs probabilities and we cannot evaluate a binary classification without first setting a threshold. Though we report multiple metrics, we find it valuable to use only one metric for our decision criterion.

V. Evaluation

We attempt to find a method of ranking features together with a model that yields predictive performance across all partitions, which currently represent different phases of the solar cycle in our benchmark dataset. We do this by computing two feature rankings and evaluating the performance of a model trained on these features against all partitions. The final result should demonstrate performance with a good TSS and HSS measure.

Skill scores, such as TSS and HSS, are prominent in rare event forecasting, and ideally they should be evaluated both independently and together to understand the model behavior.

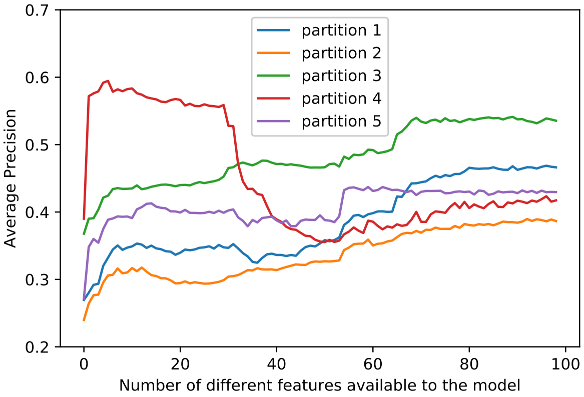

Fig. 4. Model performance in terms of Average Precision with varying number of features selected based on F-score.

However, we need one score for the model training and use average precision as our performance evaluation metric. This is because, average precision provides a balanced performance evaluation between TSS, which focuses on improving the number of true predictions, and HSS, which also considers the improved skill over standard random prediction of rare and more common events. Additionally, using average precision we can evaluate our model's performance across all classification thresholds.

We use RF with the maximum depth set to twelve, we allow all features to be considered at each node, and we use 1000 estimators.

To determine whether our feature ranking aids model performance, we evaluate using only the top 100 features from each ranking. Because each of our 1032 features is a statistical feature of a time series of a physical parameter, we expect most features to be redundant. A successful feature ranking should filter out those which yield lower performance.

To evaluate performance, we use the same cross-validation described in Section IV-C. We train a model using only the top 100 features and test the performance measured by average precision. We then remove the lowest ranked feature, and repeat the process. The result is depicted in Figure 4 for features ranked by F-score and Figure 3 for features ranked by MDI.

On two partitions, we lose a bit of performance, but on three partitions we gain performance. We summarize the original performance of the model when all 1032 features in Table III and the performance when only features selected based on MDI are used in Table IV.

To demonstrate the relationship between model performance determined by average precision, and the more familiar TSS and HSS for binary classification, we include Figure 5. We select a classification threshold based on the maximum value of HSS and report the TSS for the same threshold. These values for all partitions are listed in Table II

We believe this method is suitable for evaluating the relationship between the number of features used and model performance, because it allows us to examine changes in behavior across different periods of the solar cycle while still confirming that we are achieving a good predictive performance.

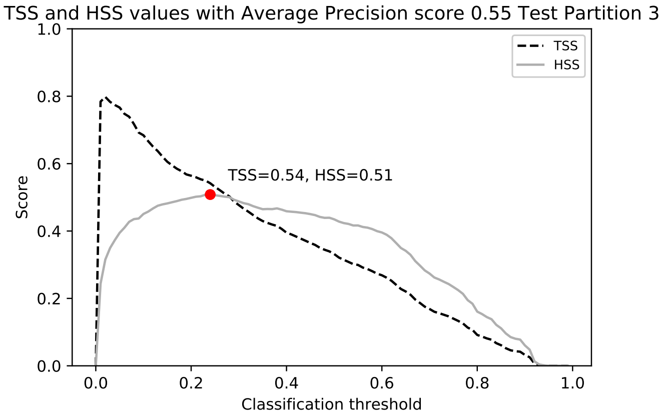

Fig. 5. TSS and HSS values for each threshold, given AP

TABLE II
HIGHEST AVERAGE PRECISION ACHIEVED AGAINST EACH PARTITION AND THE TSS AND HSS RESULTING FROM A SELECTED BINARY CLASSIFICATION

Partition	Average Precision	TSS	HSS
Partition 1	0.47	0.5	0.44
Partition 2	0.43	0.54	0.43
Partition 3	0.53	0.54	0.51
Partition 4	0.52	0.42	0.5
Partition 5	0.45	0.51	0.44

VI. Conclusion

We attempted to find a feature ranking which allows us to determine the minimum number of features necessary to demonstrate high predictive performance on a new time series benchmark dataset designed for solar flare forecasting. We find that we are able to reduce the dimensionality of our dataset significantly, and still maintain good predictive performance against all partitions, representing different phases of the solar cycle. Using average precision as our metric for ranking model

TABLE III
PREDICTIVE PERFORMANCE AGAINST EACH PARTITION WHEN ALL 1032 FEATURES USED FOR TRAINING

Partition	Average Precision
Partition 1	0.47
Partition 2	0.43
Partition 3	0.53
Partition 4	0.52
Partition 5	0.45

TABLE IV
PREDICTIVE PERFORMANCE AGAINST EACH PARTITION WHEN ONLY SELECTED FEATURES USED FOR TRAINING

Partition	Average Precision	Number of Features Used
Partition 1	0.47	96
Partition 2	0.43	92
Partition 3	0.53	89
Partition 4	0.52	5
Partition 5	0.45	59

performance, we achieve a balance of TSS and HSS that indicate a robust forecasting method. We discuss some of the difficulties and pitfalls of a classification problem with severe class imbalance, and demonstrate some methods used to handle the problem.

In the future, we plan to continue working toward finding specific features that may have significant importance in solar flare activity. We also plan to reduce the granularity of the data by testing different methods for discretizing the continuous features. So far we have investigated the entire dataset, which represents almost eight years of solar activity, with the aim of finding a generic forecast for predicting flares during any phase of the solar cycle. We will consider working toward achieving good predictive performance and interpretable results across all partitions, but we also plan to investigate individual partitions more thoroughly.

ACKNOWLEDGMENT

This work was supported in part by two NASA Grant Award [No. NNH14ZDA001N], and one NSF Grant Awards [No. AC1443061 and AC1931555]. The AC1443061 award has been supported by funding from the Division of Advanced Cyber infrastructure within the Directorate for Computer and Information Science and Engineering, the Division of Astronomical Sciences within the Directorate for Mathematical and Physical Sciences, and the Division of Atmospheric and Geospace Sciences within the Directorate for Geosciences.

REFERENCES

[1] S. S. Board, N. R. Council *et al.*, *Severe space weather events: Understanding societal and economic impacts: A workshop report*. National Academies Press, 2009.

[2] T. Maynard, N. Smith, and S. Gonzalez, "Solar storm risk to the north american electric grid," *Lloyd's*, vol. 1, p. 11, 2013.

[3] M. H. MacAlester and W. Murtagh, "Extreme space weather impact: An emergency management perspective," *Space Weather*, vol. 12, no. 8, pp. 530–537, 2014.

[4] E. K. Fry, "The risks and impacts of space weather: Policy recommendations and initiatives," *Space Policy*, vol. 28, no. 3, pp. 180–184, 2012.

[5] C. Schrijver, J. Beer, U. Baltensperger, E. Cliver, M. Güdel, H. Hudson, K. McCracken, R. Osten, T. Peter, D. Soderblom *et al.*, "Estimating the frequency of extremely energetic solar events, based on solar, stellar, lunar, and terrestrial records," *Journal of Geophysical Research: Space Physics*, vol. 117, no. A8, 2012.

[6] M. Shea, D. Smart, K. McCracken, G. Dreschhoff, and H. E. Spence, "Solar proton events for 450 years: The carrington event in perspective," *Advances in Space Research*, vol. 38, no. 2, pp. 232–238, 2006.

[7] M. Hüsken and P. Stagge, "Recurrent neural networks for time series classification," *Neurocomputing*, vol. 50, pp. 223–235, 2003.

[8] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "Lstm fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2017.

[9] C. J. Schrijver, "A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting," *The Astrophysical Journal Letters*, vol. 655, no. 2, p. L117, 2007.

[10] M. G. Bobra and S. Couvidat, "Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm," *The Astrophysical Journal*, vol. 798, no. 2, p. 135, 2015.

[11] J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat, M. C. Rabello-Soares, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, D. J. Akin, B. A. Allard, J. W. Miles, R. Rairden, R. A. Shine, T. D. Tarbell, A. M. Title, C. J. Wolfson, D. F. Elmore, A. A. Norton, and S. Tomczyk, "Design and Ground Calibration of the Helioseismic and Magnetic Imager (HMI) Instrument on the Solar Dynamics Observatory (SDO)," *solphys*, vol. 275, no. 1-2, pp. 229–259, Jan 2012.

[12] C. Campi, F. Benvenuto, A. M. Massone, D. S. Bloomfield, M. K. Georgoulis, and M. Piana, "Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence," *Unpublished*, 2019.

[13] K. Florios, I. Kontogiannis, S.-H. Park, J. A. Guerra, F. Benvenuto, D. S. Bloomfield, and M. K. Georgoulis, "Forecasting solar flares using magnetogram-based predictors and machine learning," *Solar Physics*, vol. 293, no. 2, p. 28, 2018.

[14] M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. Leka, "The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: Sharps–space-weather hmi active region patches," *Solar Physics*, vol. 289, no. 9, pp. 3549–3578, 2014.

[15] K. Domijan, D. S. Bloomfield, and F. Pitié, "Solar flare forecasting from magnetic feature properties generated by the solar monitor active region tracker," *Solar Physics*, vol. 294, no. 1, p. 6, 2019.

[16] P. A. Higgins, P. T. Gallagher, R. J. McAteer, and D. S. Bloomfield, "Solar magnetic feature detection and tracking for space weather monitoring," *Advances in Space Research*, vol. 47, no. 12, pp. 2105–2117, 2011.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[19] ——, *Classification and regression trees*. Routledge, 2017.

[20] B. A. M. K. G. D. J. K. S. M. A. Ahmadzadeh, M. Hostetter and R. A. Angryk, "Challenges with extreme class-imbalance and temporal coherence: A study on solar flare data," in *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, in press.

[21] G. Louppe, "Understanding random forests: From theory to practice," *arXiv preprint arXiv:1407.7502*, 2014.

[22] D. S. Bloomfield, P. A. Higgins, R. J. McAteer, and P. T. Gallagher, "Toward reliable benchmarking of solar flare forecasting methods," *The Astrophysical Journal Letters*, vol. 747, no. 2, p. L41, 2012.