Solar Phys (2019) 294:6
https://doi.org/10.1007/s11207-018-1392-4

Solar Flare Forecasting from Magnetic Feature Properties Generated by the Solar Monitor Active Region Tracker

Katarina Domijan[1] · **D. Shaun Bloomfield**[2] · **François Pitié**[3]

Received: 17 September 2018 / Accepted: 31 December 2018

Abstract We study the predictive capabilities of magnetic-feature properties (MF) generated by the Solar Monitor Active Region Tracker (SMART: Higgins *et al.* in *Adv. Space Res.* **47**, 2105, 2011) for solar-flare forecasting from two datasets: the full dataset of SMART detections from 1996 to 2010 which has been previously studied by Ahmed *et al.* (*Solar Phys.* **283**, 157, 2013) and a subset of that dataset that only includes detections that are NOAA active regions (ARs). The main contributions of this work are: we use marginal relevance as a filter feature selection method to identify the most useful SMART MF properties for separating flaring from non-flaring detections and logistic regression to derive classification rules to predict future observations. For comparison, we employ a Random Forest, Support Vector Machine, and a set of Deep Neural Network models, as well as lasso for feature selection. Using the linear model with three features we obtain significantly better results (True Skill Score: TSS = 0.84) than those reported by Ahmed *et al.* (*Solar Phys.* **283**, 157, 2013) for the full dataset of SMART detections. The same model produced competitive results (TSS = 0.67) for the dataset of SMART detections that are NOAA ARs, which can be compared to a broader section of flare-forecasting literature. We show that more complex models are not required for this data.

Keywords Flares, Forecasting · Flares, Relation to Magnetic Field · Active Regions, Magnetic Fields

1. Introduction

Solar flares strongly influence space weather, and their prediction using photospheric magnetic field observations has been studied extensively in recent years, *e.g.* Abramenko (2005),

✉ K. Domijan
katarina.domijan@mu.ie

1 Department of Mathematics and Statistics, Maynooth University, Maynooth, Co. Kildare, Ireland

2 Department of Mathematics, Physics and Electrical Engineering, Northumbria University, Newcastle Upon Tyne, NE1 8ST, UK

3 Department of Electronic and Electrical Engineering, Trinity College Dublin, Dublin, Ireland

McAteer, Gallagher, and Ireland (2005), Schrijver (2007), Leka and Barnes (2007), Georgoulis and Rust (2007), Qahwaji *et al.* (2008), Colak and Qahwaji (2008, 2009), Barnes and Leka (2008), Mason and Hoeksema (2010), Yu *et al.* (2010), Yang *et al.* (2013), Al-Ghraibah, Boucheron, and McAteer (2015), Boucheron, Al-Ghraibah, and McAteer (2015), Bobra and Couvidat (2015), Liu *et al.* (2017b, 2017a), Daei, Safari, and Dadashi (2017), Gheibi, Safari, and Javaherian (2017), Raboonik *et al.* (2017), Nishizuka *et al.* (2017, 2018), and Huang *et al.* (2018).

In this article we analyzed a dataset of magnetic feature (MF) properties generated by the Solar Monitor Active Region Tracker (SMART: Higgins *et al.*, 2011), an automated system for detecting and tracking active regions (AR) from the *Solar and Heliospheric Observatory* (SOHO) *Michelson Doppler Imager* (MDI: Scherrer *et al.*, 1995) magnetograms. SMART determines MF properties such as region size, total flux, flux imbalance, flux emergence rate, Schrijver's $\mathcal{R}$-value, and Falconer's measurement of non-potentiality. Each MF detection was classed as flaring or non-flaring if it produced a C-class or above flare within the 24 hours following the observation.

This dataset was previously analyzed by Ahmed *et al.* (2013) and in this article one of the aims was to improve on their results. We considered a number of classification approaches: binary logistic regression (LR: Cox, 1958), which is a linear classifier, as well as the classifiers that allow for nonlinear classification rules, namely Random Forests (RFs: Breiman, 2001), support vector machine (SVM: Vapnik, 1998), and a set of Deep Feedforward Neural Network (DNN) architectures. We considered feature selection for the linear model: the LR classifier was applied to a small subset of MF properties selected by a marginal relevance (MR) criterion (Dudoit, Fridlyand, and Speed, 2002) and the full feature set. We also used lasso (Tibshirani, 1996), a model related to LR that simultaneously performs classification and feature selection.

To assess how the results of a predictive model will generalize to an independent dataset, we used cross-validation where the training of algorithms and feature selection are carried out on the training set and the presented results are shown for the test set. True Skill Scores (TSS), Heidke Skill Score (HSS), Receiver Operating Characteristic (ROC) curves, and Area Under ROC curve (AUC) were used as measures of classifier performance.

For the dataset analyzed by Ahmed *et al.* (2013) we found that the linear classifier using only the top three features selected by MR yielded good classification rates with the highest TSS of 0.84, sensitivity (recall) of 95 %, and specificity of 89 %. This is a significant improvement on the previous analysis of this data. None of the other approaches that we considered exceeded this performance.

SMART detects MFs automatically and independently from NOAA active regions. A large number of detections are small magnetic-flux regions that have no associated sunspot structure and do not possess many of the properties that SMART calculates, yielding values close to zero for some of the features. These detections never flare, and it is relatively easy for a forecasting system to get them correct. In order to compare our results to a broader section of the flare-forecasting literature, we analyzed a second set of results that correspond to SMART detections that are NOAA active regions by initially filtering the SMART dataset. For this reduced dataset, the same linear classifier with the top three features selected by MR yielded TSS of 0.67 with corresponding sensitivity and specificity of 87 % and 80 %. None of the other models, including more comprehensive searches of the feature space and nonlinear classifiers, were able to improve on this performance.

Based on the classification results as well as the visualization of the data, we show that there is no advantage in including a larger number of features or fitting more complex, nonlinear models for these datasets.

For comparison with Ahmed *et al.* (2013), we used the same split of the data into training and testing sets. For the full dataset of all SMART detections, the training set is large, comprising 330,000 instances and as such puts constraints on the choice of classifier methodology. For example, kernel-based classifiers such as SVM require the computation of an $n \times n$ dimensional kernel matrix and do not scale well to data where n, the number of instances, is large. To evade this issue we took the approach of subsampling from the full training set to construct 50 smaller training sets of 400 instances each. The SVMs were trained on these small training sets whereas DNNs involve highly parameterized models and were trained on the full training set. LR and RF were trained on the small subsampled training sets and the full training set. We found that using the entire training set to train the algorithms gives no improvement in test classification rates.

The analysis of the data was done in R, a free software environment for statistical computing and graphics (R Core Team, 2017). The graphical displays were produced using the ggplot2 (Wickham, 2009) and plot3D (Soetaert, 2017) packages. The code to reproduce the analysis and graphics can be accessed at github.com/domijan/Sola.

The article is organized as follows: in Section 2 we describe the dataset, in Section 3 we briefly outline the method used for feature selection, the classification algorithms, the cross-validation settings for assessing classifier performance, and the forecast performance measures. Section 4 presents the results, and in Section 5 we make some concluding comments.

2. Data

Data are line-of-sight magnetograms from SOHO/MDI. Magnetic-feature properties were extracted by the Solar Monitor Active Region Tracking algorithm (Higgins *et al.*, 2011). Flares are from *Geostationary Operational Environmental Satellite* (GOES) soft X-ray (1 – 8 Å) flare lists provided by NOAA/SWPC.

SMART detects MFs automatically and independently from NOAA active regions. Following Ahmed *et al.* (2013) we defined an "MF detection" as an individual SMART MF detected in one MDI magnetogram. Each MF detection was classified as flaring or non-flaring according to whether it produced a C-class or above flare within the 24 hours following the observation. In order to minimize the error caused by projection effects, only MF detections located within 45 degrees from solar disc center were considered. The dataset comprises MF detections generated by SMART from April 1996 – December 2010. A list of SMART MF features used in this analysis with descriptions is given in Table 8.

In this article we study two datasets: the "full SMART dataset" of all MF detections generated by SMART from 1 April 1996 to 31 December 2010 and a "NOAA AR dataset" containing only those SMART MF detections that can be associated with NOAA ARs. The second dataset is derived by retaining only those SMART MF detections whose boundaries encompass the coordinates of one or more NOAA ARs after these were time-rotated to the MDI observation times used by SMART.

3. Methods

3.1. Classification Algorithms

Logistic regression is a well established framework for modeling and prediction of data where the response variable of interest is binary. It is a subset of the Generalized Linear

Models (GLMs: Nelder and Wedderburn, 1972), which are widely used across a range of scientific disciplines and are available in almost all statistical software packages.

For each MF detection in a training dataset, we have a feature vector $[\boldsymbol{X}_i]$ and an observed class label $[Y_i \in \{0, 1\}]$, denoting if it produced a flare. The distribution of Y_i is modeled by a Bernoulli $[p_i]$ distribution, where $p_i = P(Y_i = 1|\boldsymbol{X}_i, \boldsymbol{\beta})$ denotes the probability of flare and $\boldsymbol{\beta}$ is the parameter vector. In this model we use the logit link, where p_i is the logistic function of a linear combination of the explanatory features:

$$p_i = \frac{1}{1 + e^{\beta_0 + \boldsymbol{\beta} \boldsymbol{X}_i}}.$$

The model coefficients $\boldsymbol{\beta}$ are estimated using the maximum-likelihood method and are used to estimate p_i. The class of a MF detection i can be predicted by thresholding the estimated p_i at a particular value, therefore giving a linear classification algorithm. As such, LR is a related model to Fisher's linear discriminant analysis (LDA: Fisher, 1936) and SVMs with linear kernels. Other classification methods that we considered were chosen because they take very different approaches to inducing nonlinearity, feature selection, and model fitting.

LR is an example of a feedforward-neural-network architecture with a single neuronal unit, single layer, and sigmoid activation function. By adding units and layers, the neural networks extend the LR model to complex models with nonlinear classification boundaries. The architectures with two or more hidden layers are generally called *deep neural networks* (DNNs: Géron, 2018). There are many types of DNN architectures and these models that have been successfully applied in the domains of text and image analysis. In this article we employed a multi-layer perceptron (MLP) that consists of a sequence of densely connected layers of neurons. This is the classic architecture for the data where the feature vector does not have a hierarchical structure, as is the case for image or text data.

In this article we considered a range of fully connected layers architectures: two hidden layers with 8 and 4 units (DNN_8_4), two hidden layers with 16 units each (DNN_16_16), two hidden layers with 256 and 32 units (DNN_256_32) and three hidden layers with 13, 6, and 6 units each (DNN_13_6_6). We chose tanh activations for the hidden units. The output layer for each of the networks is a single sigmoid unit and the loss function is set as the binary cross-entropy. The networks have been trained over 200 epochs with a mini-batch size of 1024 and using the adaptive moment estimation (ADAM) optimization strategy. We did consider deeper architectures, but they were overfitting the data; and we also considered different activation functions, but there was no difference in the algorithm performance.

A *support vector machine* (SVM) is a kernel extension of a binary linear classifier that constructs a hyperplane to separate two classes. The hyperplane is chosen so that the smallest perpendicular distance of the training data to the hyperplane (margin) is maximized. A tuning parameter (cost) controls the number of observations that are allowed to violate the margin or the hyperplane. The kernel trick is a general technique that can be applied to any optimization problem that can be rewritten so that it takes the inner products between pairs of the training observations as opposed to the observations themselves. When the inner product is replaced with a more general kernel, the observations are implicitly mapped to a higher-dimensional feature space where the optimization takes place. For linear classifiers, this has the effect of fitting nonlinear decision boundaries in the original feature space. The shape of the boundary is determined by the choice of the kernel and its parameterization.

Random Forests (RFs) provide a different approach to the classification problem and to feature selection. They grow a number of decision trees on bootstrapped samples of the training set. Each tree recursively partitions the feature space into rectangular subregions

where the predicted class is the most common occurring class. At each iteration, a tree algorithm searches through all of the possible split-points along a randomly selected subset of features to find a partition which minimizes the region impurity, measured by the Gini index. For a binary problem, the Gini index is given by $2\hat{p}_m(1-\hat{p}_m)$, where $\hat{p}_m$ is the proportion of flaring observations in region m. A single consensus prediction is obtained from all the trees using a majority vote, which allows for very complex and nonlinear decision boundaries. The total decrease in the Gini index from splitting on a feature, averaged over all trees, can be taken as an estimate of that feature's importance.

3.2. Feature Selection

For this study, we use the Marginal Relevance (MR) score to rank the features in order of their capability to discriminate between the two classes (flare/non-flare). The MR score for each feature is the ratio of the between-class to within-class sum of squares. This idea underpins many statistical methodologies and is frequently used in genetics to screen out a large number of spurious features; see, for example, Dudoit, Fridlyand, and Speed (2002).

The approach to feature selection using MR screens out the unnecessary features before applying logistic regression. MR considers the information in each feature independently so the highest-ranked features can be correlated and do not necessarily form the optimal subset for the purposes of classification.

For these datasets we also fitted lasso (Tibshirani, 1996), a model related to LR, but where the coefficients β are simultaneously shrunk to zero using a penalty that is controlled by a tuning parameter. Lasso provides a more sophisticated approach to feature selection than MR and simultaneously reduces dimensionality of the feature space and performs classification. Lasso is implemented in the R package glmnet (Friedman, Hastie, and Tibshirani, 2009).

All features were used to train the nonlinear classifiers. SVMs combine all of the feature information into a distance matrix (kernel) and can cope with correlated inputs and a small number of spurious features. In RF all of the features are used to grow the trees, and at each iteration randomly selected subsets are jointly considered for subdividing the feature space.

3.3. Cross-validation

For consistency and comparison with Ahmed *et al.* (2013) we use the MF detections from April 1996 – December 2000 and January 2003 – December 2008 to train the classification algorithms and the MF detections from January 2001 – December 2002 and January 2009 – December 2010 comprise the test set.

The number of flaring/non-flaring SMART detections in the training and testing sets for both the full and the NOAA AR data are shown in Table 1.

The training set is further subsampled by randomly drawing 200 instances of flares and 200 instances of non-flares to form 50 smaller training sets.

The full SMART dataset contains 490,997 non-flaring and 27,244 (5.4 %) flaring instances and therefore exhibits a large class imbalance. In the NOAA AR dataset 18.6 % of detections were classed as flares. Class imbalance is a common problem and has received a great deal of attention in the classification literature; see, for example, Chawla, Japkowicz, and Kotcz (2004). In the construction of the subsampled training sets we uniformly sampled instances of flares and non-flares but adjusted the mixture of the classes, an approach known as case-control sampling. Logistic regression models fitted to subsamples can be converted to a valid model using a simple adjustment to the intercept; see Fithian and Hastie (2014).

Table 1 The number of flaring/non-flaring SMART detections in the full and reduced datasets.

	Full SMART dataset		NOAA AR dataset	
	Training set	Testing set	Training set	Testing set
flare	16,673	10,571	1,137	707
non-flare	313,617	177,380	5,272	2,789

3.4. Forecast Performance Measures

In a binary classification problem we can designate one outcome as positive (flare) and the other as negative (no flare). For algorithms with probabilistic outputs, binary forecasts are obtained by thresholding p, *e.g.* predicting a flare if the estimated $p > 0.5$. A confusion matrix is constructed by cross-tabulating the predicted with the observed classes. This presents the number of true positives TP (flare predicted and observed), false positives FP (flare predicted but not observed), true negatives TN (no flare predicted and none observed), and false negatives FN (no flare predicted but observed).

The true positive rate (TPR), or sensitivity, is the proportion of correctly classified flares out of all of the flares observed in the sample TPR = TP/ (TP + FN). The true negative rate (TNR), or specificity, is the proportion of true negatives out of all the non-flaring instances. The false positive rate is FPR = 1 − TNR, and the false negative rate is FNR = 1 − TPR.

A classifier that performs well will give a high TPR and TNR and, consequently, low FPR and FNR. For classifiers that give probabilistic outputs, the sensitivity (TPR) can be increased by lowering the threshold of p, but this automatically increases the FPR. An optimal choice of the threshold is context dependent: the cost of FNs might be higher than FPs. For a $(0, 1)$ range of thresholds, a receiver operating characteristic (ROC) curve plots the TPR *vs.* FPR. The ROC curve and the corresponding area under the ROC curve (AUC) are used for comparing the performance of algorithms over the entire range of thresholds. The ideal ROC curve is in the top-left corner, giving high TPR and a low FPR, and the maximum possible value for AUC is 1.

For a single threshold or for classifiers with non-probabilistic outputs, the elements of the confusion matrix can be combined in a number of ways to obtain a single measure of the performance of a given method.

The accuracy (ACC) gives the proportion of correctly classified observations over both classes.

The true skill statistic (TSS: Youden, 1950; Hanssen and Kuipers, 1965) combines the sensitivity and specificity by taking TSS = TPR + TNR − 1.

The Heidke skill score (HSS: Heidke, 1926) measures the fraction of correct predictions after adjusting the predictions that would be correct due to random chance.

For more details as regards forecast performance measures used in solar-flare literature see Bloomfield *et al.* (2012), Barnes and Leka (2008), Barnes *et al.* (2016).

4. Results

LR with top three features selected by MR (LR) and SVMs were trained on 50 subsampled training sets. The results for the SVM were obtained from the R-package e1071 (Meyer *et al.*, 2017), with a Gaussian kernel with the bandwidth parameter set to 0.03 for the full SMART dataset and 0.01 for the NOAA AR dataset. The cost of the constraints violation was set to 1.

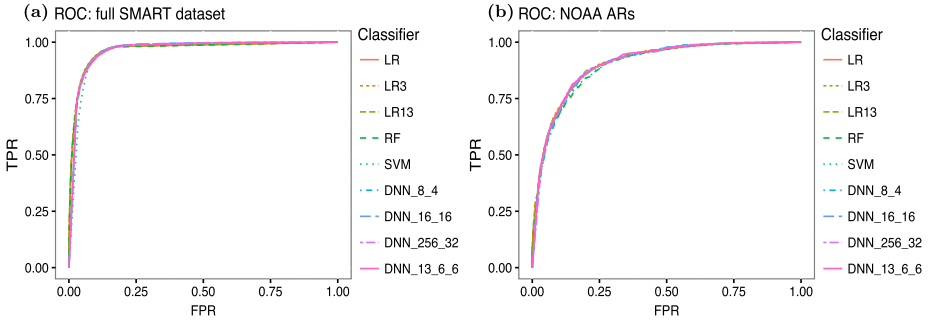

Figure 1 ROC curves, calculated for the training set for LR (LR with three features trained on subsamples), LR3 (LR with three features trained on the full training set), LR13 (LR with all features trained on the full training set), RF, SVM, and the DNN architectures. (**a**) Full SMART dataset, (**b**) NOAA ARs dataset.

LR with top three features selected by MR (LR3), the full set of features (LR13), RF and DNN were trained on the full training set. The DNN architectures were: two layers with 8 and 4 units (DNN_8_4), two layers with 16 units each (DNN_16_16), two layers, 256 and 32 units (DNN_256_32), and three layers 13, 6 and 6 units each (DNN_13_6_6).

For RF, 500 trees were grown, where at each iteration three variables were randomly sampled as candidates for each split. The tuning parameters of RF and SVM were tuned over grids using cross-validation of the training set. For support vector machines (SVMS) we tried out Gaussian, Anova, and Laplacian kernels and Bayesian Kernel Projection Classifier (BKPC) (Domijan and Wilson, 2011), a sparse Bayesian variant using lower dimensional projections of the data in the feature space. All three kernels and BKPC performed equally well in the training set at the optimal values of their kernel parameters, so in this article we present the results of the Gaussian kernel as it is best known. RF is implemented in the R library randomForest (Liaw and Wiener, 2002). MR algorithm is implemented in the R library BKPC (Domijan, 2016). DNNs were fitted using the Keras library in R (Allaire and Chollet, 2017).

For the purposes of analysis, some of the features were log transformed (high-gradient neutral-line length in the region, neutral-line length in the region, Falconer's WL_{SG}-value, Schrijver's $\mathcal{R}$-value, total unsigned magnetic flux, and flux emergence rate). The same transformations were found to be adequate for both the full SMART dataset and the NOAA AR dataset. For both datasets, the features in the training sets were scaled to have zero mean and unit variance, and the same scaling was then applied to the test sets.

The forecast performance measures (TPR, TNR, TSS, ACC, and HSS), described in Section 3.4, were calculated for the test set at a range of thresholds of p.

For the classifiers trained on 50 subsampled training sets, 50 classification rules were obtained and consequently the median values, 2.5th and 97.5th percentiles of the resulting forecast performance measures are reported. This can be used to assess the sensitivity of the algorithms to the choice of the training sets.

4.1. Full SMART Dataset

Figure 1a shows the ROC curves plotted for all of the classifiers for the full SMART dataset. For the algorithms trained on 50 subsampled training sets (LR and SVM), the ROC curve is obtained from the median TPR and FPR over the $(0, 1)$ range of thresholds. The ROC

Table 2 AUC, the highest TSS, and HSS for all the classifiers. The first two (LR and SVM) were fitted to subsampled training sets. The rest were trained on the full training set.

Classifier	TSS	HSS	AUC
LR	0.84 (0.83, 0.84)	0.63 (0.60, 0.64)	0.966 (0.962, 0.968)
SVM	0.83 (0.83, 0.84)	0.56 (0.51, 0.60)	0.949 (0.942, 0.956)
LR3	0.84	0.64	0.967
LR13	0.84	0.64	0.967
RF	0.83	0.63	0.964
DNN_8_4	0.83	0.63	0.966
DNN_16_16	0.83	0.63	0.966
DNN_256_32	0.83	0.63	0.966
DNN_13_6_6	0.83	0.64	0.965

curves are very close and show that after careful tuning, all models perform equally well and converge to the same results in terms of the performance measures.

AUC, the highest TSS, and HSS for all classifiers for the full SMART data are given in Table 2. For algorithms trained on 50 subsamples; the reported value is the median with 2.5th and 97.5th percentiles given in brackets. AUC ranged from 0.949 to 0.967, TSS ranged from 0.83 to 0.84, and HSS ranged from 0.56 to 0.64. This again shows that no model convincingly outperformed the others in terms of predictive ability for this dataset. The linear model with only three features works as well as the more complex models that allow for nonlinear classification boundaries. Likewise, including extra features in the linear model did not improve performance. The results for LR with three features are the same for the algorithm trained on the subsampled training sets and the full training set (LR and LR3), showing that small datasets of 400 instances are sufficient to train this model. Narrow confidence bands for LR and SVM indicate that the classification results are consistent across the subsampled training sets.

For the logistic regression algorithm with the three input features trained on the subsampled training sets (LR) the median values for TSS, TPR, TNR, ACC, and HSS at each threshold are presented in Table 3. The 2.5th and 97.5th percentiles for TSS and HSS are given in brackets. For comparison, Table 4 presents the results for the same algorithm trained on the full training set (LR3).

For LR, the highest TSS of 0.84 is obtained at the thresholds between 0.04 to 0.08 which give the TPR in the range of 0.96 and 0.92 and TNR of 0.88 and 0.91, respectively. The results from Ahmed *et al.* (2013) give a TPR of 0.523, TNR of 0.989 with HSS of 0.595 for the machine-learning algorithm and TPR of 0.814 and HSS of 0.512 (TNR is not reported) for the automated solar-activity prediction (ASAP). Area under ROC curve (not reported by Ahmed *et al.*, 2013) was calculated for the 50 curves and the median AUC value for LR is 0.966 with 2.5th and 97.5th percentiles of 0.962, 0.968. For the same model trained on the full training set using three features (LR3), the AUC was 0.967. Using all features in the model (LR13) did not increase AUC from 0.967.

4.2. Choice of Skill Scores and Threshold

The LR model fits a sigmoid surface over the range of $\boldsymbol{X}$, and the decision boundary separating the two predicted classes at any threshold is linear.

Table 3 Results from the full SMART dataset for logistic regression with three features trained on the sub-sampled training sets (LR): median values for TSS, TPR, TNR, ACC, and HSS for the testing set. 2.5th and 97.5th percentiles are given in brackets. The classification rules were obtained from 50 randomly sampled training subsets of 400 instances each.

p	TSS		TPR	TNR	ACC	HSS	
0.01	0.8	(0.79, 0.82)	0.98	0.82	0.83	0.34	(0.31, 0.37)
0.02	0.82	(0.81, 0.83)	0.97	0.85	0.86	0.38	(0.35, 0.42)
0.03	0.83	(0.82, 0.84)	0.97	0.87	0.87	0.41	(0.38, 0.45)
0.04	0.84	(0.83, 0.84)	0.96	0.88	0.88	0.43	(0.4, 0.47)
0.05	0.84	(0.83, 0.84)	0.95	0.89	0.89	0.46	(0.41, 0.49)
0.06	0.84	(0.83, 0.84)	0.94	0.9	0.9	0.47	(0.43, 0.51)
0.07	0.84	(0.83, 0.84)	0.93	0.91	0.91	0.49	(0.45, 0.52)
0.08	0.84	(0.83, 0.84)	0.92	0.91	0.91	0.5	(0.46, 0.54)
0.09	0.83	(0.83, 0.84)	0.92	0.92	0.92	0.52	(0.48, 0.55)
0.1	0.83	(0.82, 0.84)	0.91	0.92	0.92	0.53	(0.49, 0.56)
0.2	0.78	(0.74, 0.82)	0.83	0.95	0.95	0.61	(0.57, 0.64)
0.3	0.69	(0.61, 0.77)	0.72	0.97	0.96	0.63	(0.6, 0.64)
0.4	0.58	(0.5, 0.71)	0.6	0.98	0.96	0.61	(0.58, 0.64)
0.5	0.49	(0.38, 0.65)	0.5	0.99	0.96	0.57	(0.5, 0.63)
0.6	0.37	(0.26, 0.57)	0.38	0.99	0.96	0.49	(0.38, 0.61)
0.7	0.26	(0.15, 0.48)	0.26	1	0.96	0.38	(0.24, 0.57)
0.8	0.14	(0.06, 0.36)	0.14	1	0.95	0.24	(0.11, 0.48)
0.9	0.04	(0, 0.19)	0.04	1	0.95	0.07	(0.01, 0.31)

Table 4 Results from the full SMART dataset for logistic regression with three features trained on the full training dataset: TSS, TPR, TNR, ACC, and HSS for the testing set.

p	TSS	TPR	TNR	ACC	HSS
0.01	0.80	0.98	0.82	0.83	0.33
0.03	0.83	0.97	0.87	0.87	0.41
0.05	0.84	0.95	0.89	0.89	0.46
0.07	0.84	0.93	0.91	0.91	0.49
0.09	0.84	0.92	0.92	0.92	0.52
0.10	0.83	0.91	0.92	0.92	0.53
0.20	0.79	0.83	0.96	0.95	0.62
0.30	0.70	0.72	0.97	0.96	0.64
0.40	0.60	0.62	0.98	0.96	0.62
0.50	0.50	0.51	0.99	0.96	0.59
0.60	0.39	0.40	0.99	0.96	0.52
0.70	0.27	0.27	1.00	0.96	0.39
0.80	0.14	0.14	1.00	0.95	0.24
0.90	0.04	0.04	1.00	0.95	0.07

Note that in this dataset a single MF is tracked though time and will be recorded multiple times throughout its lifetime. For the purpose of this analysis, all MF detections are treated as individual measurements. This can pose a problem for interpreting the probabilistic inference of LR, which is underpinned by the independence assumption: the standard error

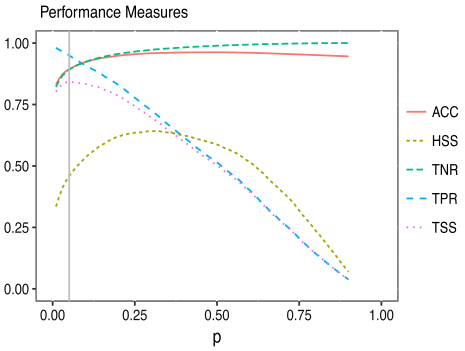

Figure 2 Sensitivity (TPR), specificity (TNR), accuracy, TSS, and HSS over a range of probability thresholds. Estimated probability was obtained from the logistic regression with three features on the full SMART dataset. The proportion of flares in the training dataset is 0.05 (*vertical line*).

estimates for the coefficients are no longer reliable and p cannot be interpreted as a probability. However, in this article, we do not make use of probabilistic inference, and we treat the logistic regression model as a deterministic linear classifier, where the choice of thresholding value of the estimated p is based on context requirements: comparing the acceptable levels of true positive and true negative rates for different thresholds. Furthermore, by training algorithms on very small subsets of the original data (200 instances of flares randomly drawn from 16,673 and 200 non-flares from 313,617) one is unlikely to get many detections of the same MF in the same subsample, which helps to evade this problem. Alternatively, one could force the randomly drawn detections to have large enough time intervals between them (*e.g.* more than two weeks) in order to ensure that the same MF is not recorded in the same training set multiple times throughout its lifetime.

Figure 2 shows the calculated sensitivity (TPR), specificity (TNR), ACC, TSS, and HSS over a range of probability thresholds (0.01 to 0.99 in steps of 0.01), where $\hat{p}$ was estimated from the LR3 model. The proportion of flaring instances in the full training set is 0.05, shown as the vertical line on the graph. This figure shows increase in TNR at the expense of TPR with increase of the threshold. At the lowest threshold $p = 0.01$, the 82 % of non-flares are correctly identified and this increases to 89 % at 0.05. Likewise, 98 % of flares are correctly predicted at the threshold of 0.01 and 95 % are correctly identified at the threshold of 0.05. The maximum estimated TSS = TNR + TPR − 1 is 0.84. Accuracy, a measure of overall error rate, is maximized at the threshold of $p = 0.5$, which gives TNR of 99 %, but this threshold misclassifies 50 % of the flares. This illustrates how ACC is a very poor choice of metric for data with a large class imbalance. At a threshold of unity, classifying all observations as negative, one will still get a 95 % accuracy score and choosing a threshold of $p = 0.5$ will misclassify over half of the flare detections. Given that the assumptions for probabilistic inference in LR are met, the p estimates the likelihood that an observation is going to flare given its feature information. Before fitting the model and utilizing the feature information, the probability that a randomly selected detection will flare is 0.05. Unlike LDA, the prior information about flare prevalence is not incorporated in the model. Thus it is sensible to take this prior as a threshold as opposed to 0.5. For this threshold, the algorithm will classify a detection as a flare if the estimated p is greater than the probability we would assign to a randomly selected detection. Figure 2 shows that this threshold strikes a sensible balance between TPR and TNR and indicates that TSS is a better choice than HSS for imbalanced data. HSS is maximized at a higher threshold of 0.3 with TPR of 0.72.

ROC curves allow for comparison of classifiers with probabilistic outputs over the range of thresholds. AUC summarizes the forecast performance in a single score, however, one could argue that comparing algorithm performance at an "optimal" threshold is more useful

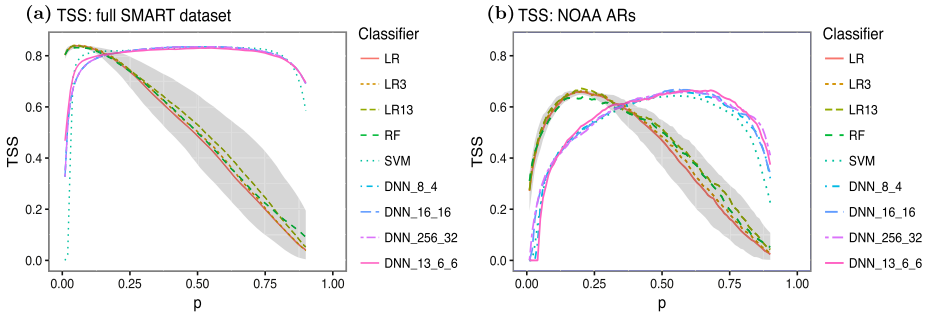

Figure 3 TSS or median TSS curve for all the classifiers. LR (LR with three features trained on subsamples), LR3 (LR with three features trained on the full training set), LR13 (LR with all features trained on the full training set), RF, SVM, and four DNN architectures. TSS curve for LR (*full line*) has 2.5th and 97.5th percentile band. (**a**) Full SMART dataset, (**b**) NOAA AR dataset.

than over the entire range. When comparing classifier performance using skill scores, we argue that it is useful to plot skill-score curves over the threshold range.

Figure 3 shows the TSS curves for different classifiers (for LR the curve is the median TSS with 2.5th and 97.5th percentile band). Figure 3a shows the TSS curves from the algorithms trained and tested on the full SMART dataset. The maximum TSS values obtained from all classifiers are very close (ranging from $0.82-0.84$) but are obtained at different thresholds for p, since the TSS curves differ in shape. This is due to the fact that the probability of a flare is estimated differently in these models. For the RF the estimate of p is non-parametric. By default, SVM produces categorical outputs; however, probabilistic extensions exist and the R implementation fits a probabilistic regression model that assumes (zero-mean) Laplace-distributed errors for the predictions. For DNNs, in order to balance the classes, the class weight of 20 was used for the flare-class labels in the computation of the loss function, which is equivalent to upsampling to match the majority class. Therefore, for the DNN models, TSS is optimized close to a threshold of 0.5.

4.3. NOAA AR Dataset

Figure 1b shows the ROC curves plotted for all the classifiers for the NOAA AR data. TSS curves are given in Figure 3b. For LR the curve is median TSS with 2.5th and 97.5th percentile band. Compared to the results of the algorithms trained and tested on the full SMART data, the performance of all algorithms is significantly weaker if trained and tested on the NOAA AR dataset, but this is still competitive with the results reported elsewhere in the solar-flare-forecasting literature; for example see Barnes *et al.* (2016).

For the NOAA AR dataset, AUC, the highest TSS, and HSS for all classifiers are given in Table 5. AUC ranged from 0.9 to 0.91, TSS ranged from 0.64 to 0.67, and HSS ranged from 0.57 to 0.59. For logistic regression algorithms (LR, LR3, LR13) and RF, the TSS is optimized at the threshold of $p \approx 0.18$, which is the prevalence of flares in the training set of these data. For DNN architectures the class weight of 5 was used for the flare class labels in the computation of the loss function. For full output of LR and LR13 results, see Table 6 and Table 7.

The results show that the algorithms with top three features (LR and LR3) perform as well as the linear model with all features (LR13) and all of the nonlinear algorithms (DNNs, RF, and SVM). In addition, small datasets of 400 instances are sufficient to train the linear

Table 5 NOAA AR data: AUC, the highest TSS, and HSS for all of the classifiers. The first two (LR and SVM) were fitted to subsampled training sets. The rest were trained on the full training set.

Classifier	TSS	HSS	AUC
LR	0.66 (0.658, 0.664)	0.59 (0.57, 0.59)	0.90 (0.90, 0.90)
SVM	0.64 (0.63, 0.66)	0.57 (0.55, 0.58)	0.90 (0.89, 0.90)
LR3	0.66	0.59	0.91
LR13	0.67	0.58	0.91
RF	0.64	0.57	0.90
DNN_8_4	0.66	0.57	0.90
DNN_16_16	0.67	0.59	0.90
DNN_256_32	0.66	0.59	0.90
DNN_13_6_6	0.66	0.58	0.90

Table 6 Results from the NOAA AR dataset for logistic regression with three features trained on the subsampled training sets (LR): median values for TSS, TPR, TNR, ACC, and HSS for the testing set. 2.5th and 97.5th percentiles are given in brackets. The classification rules were obtained from 50 randomly sampled training subsets of 400 instances each.

p	TSS		TPR	TNR	ACC	HSS	
0.1	0.59	(0.57, 0.63)	0.93	0.65	0.71	0.4	(0.37, 0.45)
0.2	0.66	(0.65, 0.66)	0.86	0.8	0.81	0.53	(0.5, 0.55)
0.3	0.63	(0.6, 0.65)	0.75	0.88	0.85	0.58	(0.56, 0.59)
0.4	0.57	(0.5, 0.61)	0.64	0.93	0.87	0.58	(0.55, 0.59)
0.5	0.47	(0.37, 0.55)	0.51	0.96	0.87	0.53	(0.45, 0.58)
0.6	0.34	(0.24, 0.47)	0.37	0.98	0.85	0.43	(0.32, 0.53)
0.7	0.22	(0.12, 0.35)	0.23	0.99	0.84	0.31	(0.18, 0.44)
0.8	0.11	(0.04, 0.24)	0.11	1	0.82	0.16	(0.07, 0.33)
0.9	0.02	(0, 0.11)	0.02	1	0.8	0.04	(0, 0.16)

Table 7 TSS, TPR, TNR, ACC, and HSS for LR13 for the NOAA AR dataset.

p	TSS	TPR	TNR	ACC	HSS
0.10	0.60	0.93	0.67	0.72	0.41
0.20	0.67	0.87	0.80	0.81	0.54
0.30	0.63	0.76	0.87	0.85	0.57
0.40	0.57	0.66	0.92	0.86	0.58
0.50	0.52	0.57	0.95	0.87	0.56
0.60	0.41	0.44	0.97	0.86	0.49
0.70	0.30	0.31	0.98	0.85	0.39
0.80	0.17	0.17	1.00	0.83	0.24
0.90	0.04	0.04	1.00	0.81	0.07

model. Narrow confidence bands for LR and SVM indicate that the classification results are consistent across the subsampled training sets.

4.4. Feature Analysis and Selection

The marginal relevance score for each feature was derived from the data used to train the classification algorithm (detections recorded from April 1996 – December 2000 and January 2003 – December 2008). Features in order of their marginal relevance derived from the full and the NOAA AR dataset are given in Table 8. The third and fourth columns give the importance order of the top six features obtained from the Random Forest in both datasets. The top features selected by lasso are given in columns five and six. Column titles with (R) denotes the importance measures were derived for the reduced NOAA AR dataset.

MR selects high-gradient neutral-line length in the region (LsgMm), maximum gradient along polarity inversion line (MxGradGpMm) and neutral-line length in the region (LnlMm) as the top three features for both full SMART and NOAA AR dataset. For both datasets, the best performing and most parsimonious lasso model had three features, but the selected area of the region (AreaMmsq) or the total un-signed magnetic flux (BfluxMx) instead of the maximum gradient along PIL (MxGradGpMm). These algorithms had the same classification performance as LR3 and LR13. In addition to neutral-line length in the region (LnlMm), features with highest importance selected by RF were Schrijver's $\mathcal{R}$-value (RvalMx), the total un-signed magnetic flux (BfluxMx), and Falconer's WL_{SG}-value (WLsgGpMm).

The forecast performance measures for the lasso models, RFs, and LR with three or more features were similar in the two datasets. Many other approaches to feature selection exist, but this indicates that there is little scope for improvement with more thorough exploration of the feature space.

Figures 4 and 5 plot the marginal densities from the training set of the full SMART dataset and NOAA AR dataset of some of the top features selected by MR, lasso, and RF. Large peaks close to zero in the non-flare distribution show that the full SMART dataset is dominated by the small magnetic-flux regions that do possess many of the properties that SMART calculates and never flare. All of the displayed features contain information on whether an observation is likely to be a flare; however, all contain a significant amount of overlap in the distributions of flaring and non-flaring regions.

Figure 6 shows one subsampled training set and test set of the full dataset of all SMART detections in three dimensions corresponding to features with the highest marginal relevance. The detections are colored by their class (flare/non-flare). In the test set, a large proportion of detections is located around zero, but due to over-plotting, it is harder to see how these observations dominate the dataset compared to density plots. These plots show that whereas the features contain information about classes, there in an overlap between them in this feature space: the classes are not perfectly separable. The shape of the classification boundary will not change this and the scope for improvement when using more complex algorithms in this feature space is limited.

5. Discussion

In order to classify MF detections we used a number of classification approaches including binary logistic regression, SVMs, RFs, and a set of DNN architectures. Categorical forecasts were obtained by thresholding the estimated probability from these models. Skill scores, curves of TSS, ROC curves, and area under the ROC were used to compare the performance of the classification approaches. We discussed the choice of skill scores and optimal thresholds for various model settings.

Table 8 SMART magnetic features in order of their marginal relevance obtained from the training dataset for the dataset of all SMART detections. The second column MR(R) is the MR feature ranking obtained from the training set of the NOAA AR detections only. Columns RF and RF(R) present the variable importance order obtained from running the Random Forest on the training set of the full and NOAA AR (R) datasets. Columns lasso and lasso (R) present the features selected in the four sparsest models fitted to the training set of the full and NOAA AR (R) datasets.

MR	MR(R)	RF	RF(R)	lasso	lasso(R)	Feature	Description
1	1	5	5	1	1	LsgMm	High-gradient neutral-line length in the region
2	3	–	6	–	–	MxGradGpMm	Maximum gradient along polarity inversion line
3	2	1	1	1	1	LnlMm	Neutral-line length in the region
4	5	4	3	–	–	WLsgGpMm	Falconer's WL_{SG}-value
5	6	6	–	1	–	AreaMmsq	Area of the region
6	7	2	2	–	–	RvalMx	Schrijver's $\mathcal{R}$-value
7	9	–	–	–	–	B	Largest magnetic field value
8	8	–	–	–	–	HGlonwdth	Heliographic longitudinal extent
9	4	3	4	–	1	BfluxMx	Total un-signed magnetic flux
10	10	–	–	–	–	HGlatwdth	Heliographic latitudinal extent
11	11	–	–	–	–	MednGrad	Median gradient along the neutral line
12	12	–	–	–	–	Bfluximb	Flux imbalance fraction in the region
13	13	–	–	–	–	DBfluxDtMx	Flux emergence rate

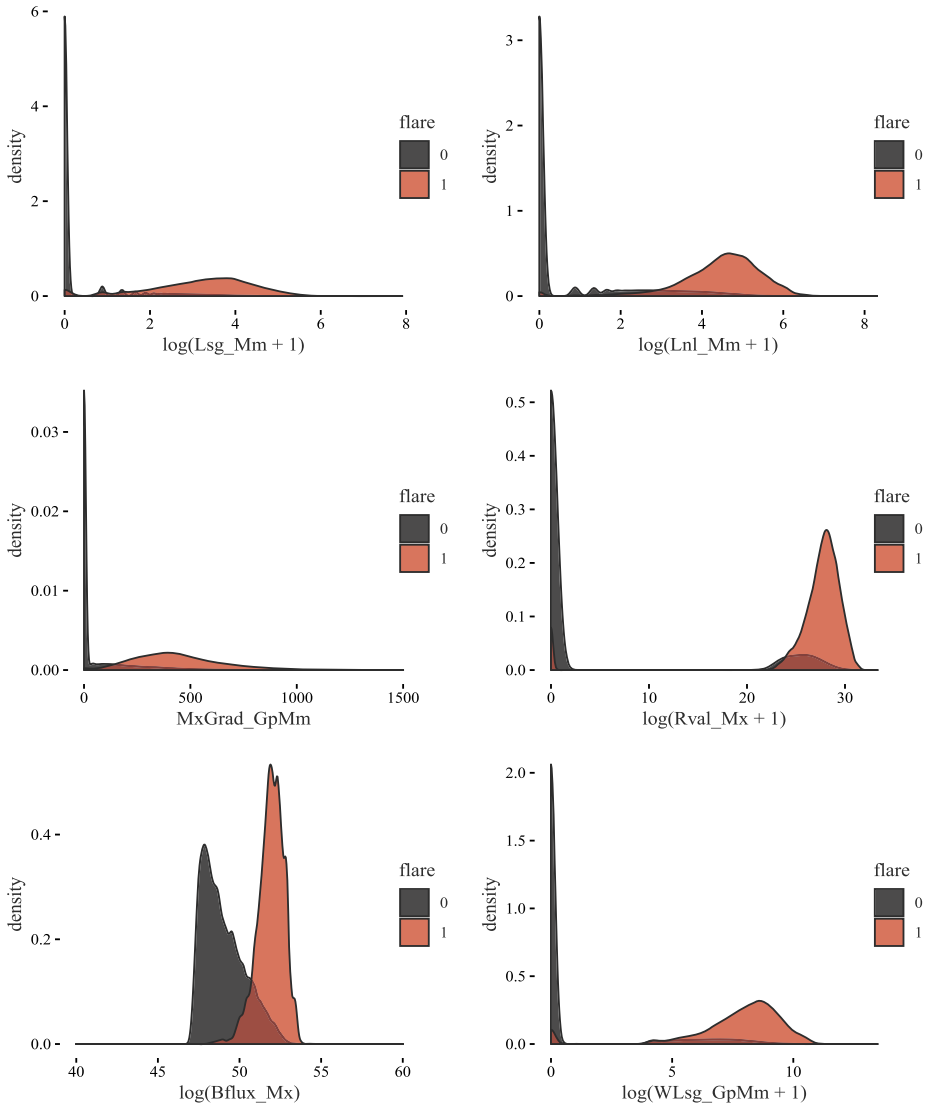

Figure 4 The marginal densities of the top features in the full SMART dataset selected by MR, lasso, and RF algorithms. The features (in order of left-to-right, top-to-bottom) are: high-gradient neutral-line length in the region (LsgMm), neutral-line length in the region (LnlMm), maximum gradient along polarity inversion line (MxGradGpMm), Schrijver's $\mathcal{R}$ value (RvalMx), total un-signed magnetic flux (BfluxMx), and Falconer's WL_{SG} value (WLsgGpMm).

The flare-prediction results that we obtained from the linear classifier with a very sparse subset of features compare favorably to those found elsewhere in the literature and show a significant improvement on the results of the previous analysis of this data. We found that, in terms of classification performance, there was no benefit in using more features or more flexible models that allow for nonlinear classification boundaries, as all approaches converged to the same result. Furthermore, we found no decrease in performance when training the algorithms on very small subsampled training sets.

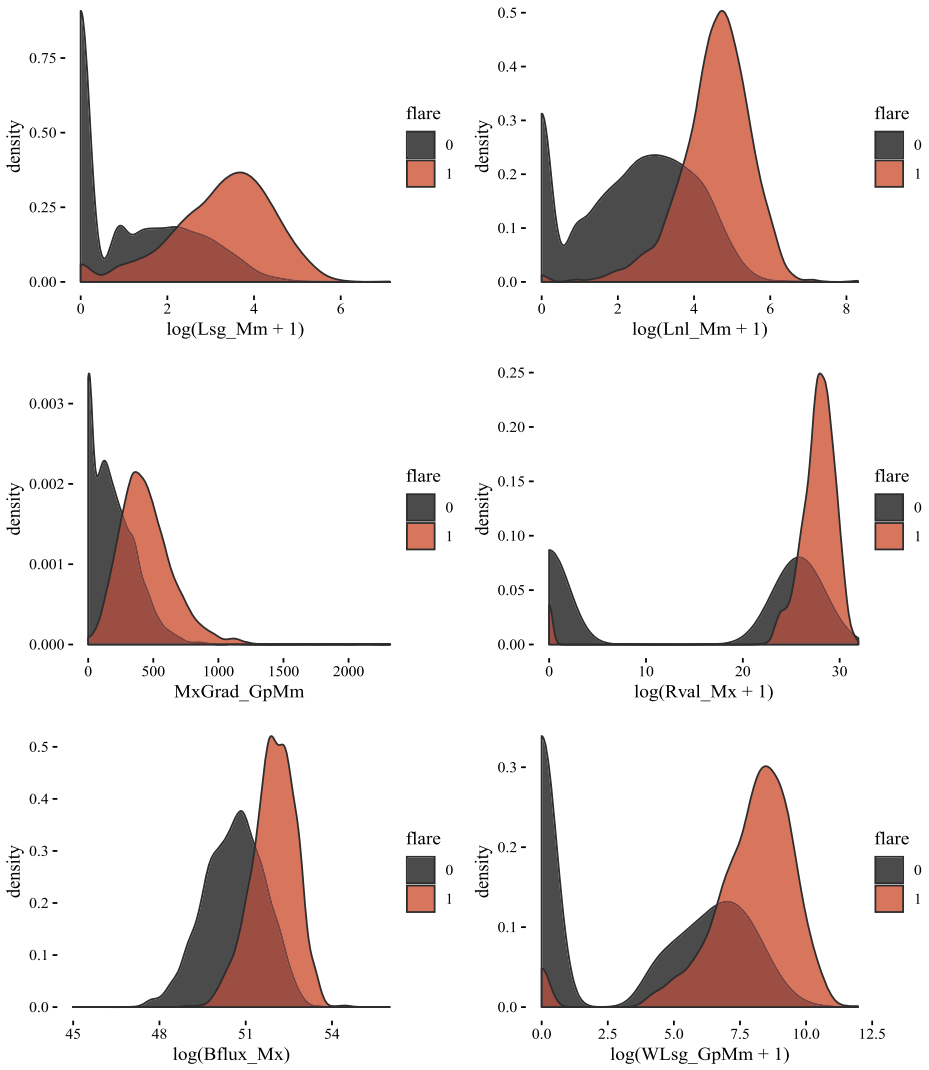

Figure 5 The marginal densities of the top features in the NOAA AR dataset selected by MR, lasso, and RF algorithms. The features (in order of left-to-right, top-to-bottom) are: high-gradient neutral-line length in the region (LsgMm), neutral-line length in the region (LnlMm), maximum gradient along polarity inversion line (MxGradGpMm), Schrijver's $\mathcal{R}$ value (RvalMx), total un-signed magnetic flux (BfluxMx), and Falconer's WL_{SG} value (WLsgGpMm).

By plotting the data in the selected dimensions we see that the classes are not perfectly separable in the space of SMART features and that there is limited scope for improvement in using more complex algorithms on this dataset.

A better performance, however, might be obtained by using the deep learning networks to learn the forecasting patterns directly from magnetograms of solar active regions as opposed to using the features computed from the magnetograms. Some work on DNNs for solar-flare prediction has been done by Nishizuka *et al.* (2018) and Huang *et al.* (2018).

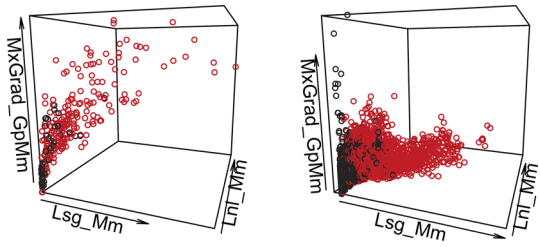

Figure 6 MF detections in the (**a**) one training dataset and (**b**) testing dataset, in three dimensions (three features with the highest marginal relevance), *red* = flare, *black* = no flare.

Direct comparisons with other published methods are difficult because of differences in the datasets, the definition of an event, and evaluation and reporting of classification results (Barnes *et al.*, 2016). It would be of interest to carry out a comparative study of classification algorithms, such as presented here, to the Space-weather HMI Active Region Patch (SHARP: Bobra *et al.*, 2014) data from the *Solar Dynamics Observatory* (SDO)/*Helioseismic and Magnetic Imager* (HMI). The data have previously been analyzed by Bobra and Couvidat (2015) and Liu *et al.* (2017a), who used SVM and RF models.

The work presented in this article is fully reproducible with code for variable selection, subsampling, and classification available via GitHub.

Disclosure of Potential Conflicts of Interest The authors declare that they have no conflicts of interest.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Abramenko, V.I.: 2005, Relationship between magnetic power spectrum and flare productivity in solar active regions. *Astrophys. J.* **629**, 1141.

Ahmed, O.W., Qahwaji, R., Colak, T., Higgins, P.A., Gallagher, P.T., Bloomfield, D.S.: 2013, Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Solar Phys.* **283**, 157. DOI.

Al-Ghraibah, A., Boucheron, L.E., McAteer, R.T.J.: 2015, An automated classification approach to ranking photospheric proxies of magnetic energy build-up. *Astron. Astrophys.* **579**, A64.

Allaire, J., Chollet, F.: 2017, *Keras: R Interface to 'Keras'*. R package version 2.1.5.9002.

Barnes, G., Leka, K.D.: 2008, Evaluating the performance of solar flare forecasting methods. *Astrophys. J. Lett.* **688**, L107. DOI.

Barnes, G., Leka, K.D., Schrijver, C.J., Colak, T., Qahwaji, R., Ashamari, O.W., Yuan, Y., Zhang, J., McAteer, R.T.J., Bloomfield, D.S., Higgins, P.A., Gallagher, P.T., Falconer, D.A., Georgoulis, M.K., Wheatland, M.S., Balch, C., Dunn, T., Wagner, E.L.: 2016, A comparison of flare forecasting methods. I. Results from the "All-Clear" workshop. *Astrophys. J.* **829**(2), 89. DOI.

Bloomfield, D.S., Higgins, P.A., McAteer, R.T.J., Gallagher, P.T.: 2012, Toward reliable benchmarking of solar flare forecasting methods. *Astrophys. J. Lett.* **747**, 2.

Bobra, M.G., Couvidat, S.: 2015, Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *Astrophys. J.* **798**(2), 135. DOI.

Bobra, M.G., Sun, X., Hoeksema, J.T., Turmon, M., Liu, Y., Hayashi, K., Barnes, G., Leka, K.D.: 2014, The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs – space-weather HMI active region patches. *Solar Phys.* **289**(9), 3549. DOI.

Boucheron, L.E., Al-Ghraibah, A., McAteer, R.T.J.: 2015, Prediction of solar flare size and time-to-flare using support vector machine regression. *Astrophys. J.* **812**(1), 51.

Breiman, L.: 2001, Random forests. *Mach. Learn.* **45**(1), 5.

Chawla, N.V., Japkowicz, N., Kotcz, A.: 2004, Editorial: special issue on learning from imbalanced data sets. *SIGKDD Explor. Newsl.* **6**(1), 1. DOI.

Colak, T., Qahwaji, R.: 2008, Automated McIntosh-based classification of sunspot groups using MDI images. *Solar Phys.* **248**, 277. DOI.

Colak, T., Qahwaji, R.: 2009, Automated solar activity prediction: a hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space Weather* **7**, S06001. DOI.
Cox, D.R.: 1958, The regression analysis of binary sequences (with discussion). *J. Roy. Stat. Soc. B* **20**, 215.
Daei, F., Safari, H., Dadashi, N.: 2017, Complex network for solar active regions. *Astrophys. J.* **845**(1), 36.
Domijan, K.: 2016, *BKPC: Bayesian Kernel Projection Classifier*. R package version 1.0.
Domijan, K., Wilson, S.P.: 2011, Bayesian kernel projections for classification of high dimensional data. *Stat. Comput.* **21**(2), 203.
Dudoit, S., Fridlyand, J., Speed, T.P.: 2002, Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**(457), 77.
Fisher, R.A.: 1936, The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179.
Fithian, W., Hastie, T.: 2014, Local case-control sampling: efficient subsampling in imbalanced data sets. *Ann. Stat.* **42**(5), 1693. DOI.
Friedman, J., Hastie, T., Tibshirani, R.: 2009, *GLMNET: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 1.1-4.
Georgoulis, M.K., Rust, D.M.: 2007, Quantitative forecasting of major solar flares. *Astrophys. J. Lett.* **661**(1), L109.
Géron, A.: 2018, *Neural Networks and Deep Learning*, O'Reilly Media, Inc, Sebastopol, CA, USA.
Gheibi, A., Safari, H., Javaherian, M.: 2017, The solar flare complex network. *Astrophys. J.* **847**(2), 115.
Hanssen, A.W., Kuipers, W.J.A.: 1965, *On the Relationship Between the Frequency of Rain and Various Meteorological Parameters: (with Reference to the Problem of Objective Forecasting)*, *Koninkl. Nederlands Meterologisch Institut. Mededelingen en Verhandelingen* **81**, Staatsdrukerij, Netherlands.
Heidke, P.: 1926, Berechnung des erfolges und der güte der windstärkevorhersagen im sturmwarnungsdienst. *Geogr. Ann.* **8**, 301. DOI.
Higgins, P.A., Gallagher, P.T., McAteer, R.T.J., Bloomfield, D.S.: 2011, Solar magnetic feature detection and tracking for space weather monitoring. *Adv. Space Res.* **47**, 2105.
Huang, X., Wang, H., Xu, L., Liu, J., Li, R., Dai, X.: 2018, Deep learning based solar flare forecasting model. I. results for line-of-sight magnetograms. *Astrophys. J.* **856**(1), 7.
Leka, K.D., Barnes, G.: 2007, Photospheric magnetic field properties of flaring versus flare-quiet active regions. IV. A statistically significant sample. *Astrophys. J.* **656**(2), 1173.
Liaw, A., Wiener, M.: 2002, Classification and regression by randomForest. *R News* **2**(3), 18.
Liu, C., Deng, N., Wang, J.T.L., Wang, H.: 2017a, Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm. *Astrophys. J.* **843**(2), 104.
Liu, J.-F., Li, F., Zhang, H.-P., Yu, D.-R.: 2017b, Short-term solar flare prediction using image-case-based reasoning. *Res. Astron. Astrophys.* **17**(11), 116.
Mason, J.P., Hoeksema, J.T.: 2010, Testing automated solar flare forecasting with 13 years of Michelson Doppler Imager magnetograms. *Astrophys. J.* **723**(1), 634.
McAteer, R.T.J., Gallagher, P.T., Ireland, J.: 2005, Statistics of active region complexity: a large-scale fractal dimension survey. *Astrophys. J.* **631**, 628. DOI.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: 2017, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-8.
Nelder, J.A., Wedderburn, R.W.M.: 1972, Generalized linear models. *J. Roy. Stat. Soc., Ser. A-G* **135**, 370.
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M.: 2017, Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. *Astrophys. J.* **835**(2), 156. DOI.
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Ishii, M.: 2018, Deep flare net (DeFN) model for solar flare prediction. *Astrophys. J.* **858**(2), 113.
Qahwaji, R., Colak, T., Al-Omari, M., Ipson, S.: 2008, Automated prediction of CMEs using machine learning of CME – flare associations. *Solar Phys.* **248**, 471. DOI.
R Core Team: 2017, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
Raboonik, A., Safari, H., Alipour, N., Wheatland, M.S.: 2017, Prediction of solar flares using unique signatures of magnetic field images. *Astrophys. J.* **834**(1), 11.
Scherrer, P.H., Bogart, R.S., Bush, R.I., Hoeksema, J.T., Kosovichev, A.G., Schou, J., Rosenberg, W., Springer, L., Tarbell, T.D., Title, A., Wolfson, C.J., Zayer, I., Team, M.E.: 1995, The solar oscillations investigation – Michelson Doppler Imager. *Solar Phys.* **162**(1-2), 129. DOI.
Schrijver, C.J.: 2007, A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting. *Astrophys. J. Lett.* **655**(2), L117.
Soetaert, K.: 2017, *plot3d: Plotting Multi-dimensional Data*. R package version 1.1.1.
Tibshirani, R.: 1996, Regression shrinkage and selection via the Lasso. *J. Roy. Stat. Soc. B* **58**, 267.
Vapnik, V.: 1998, *Statistical Learning Theory*, Wiley-Interscience, New York.

Wickham, H.: 2009, *ggplot2: Elegant Graphics for Data Analysis*, Springer, New York. 978-0-387-98140-6.
Yang, X., Lin, G., Zhang, H., Mao, X.: 2013, Magnetic nonpotentiality in photospheric active regions as a predictor of solar flares. *Astrophys. J. Lett.* **774**(2), L27.
Youden, W.J.: 1950, Index for rating diagnostic tests. *Cancer* **3**, 32.
Yu, D., Huang, X., Wang, H., Cui, Y., Hu, Q., Zhou, R.: 2010, Short-term solar flare level prediction using a Bayesian network approach. *Astrophys. J.* **710**(1), 869.