A&A 579, A64 (2015)
DOI: 10.1051/0004-6361/201525978

Astronomy & Astrophysics

An automated classification approach to ranking photospheric proxies of magnetic energy build-up

A. Al-Ghraibah[1], L. E. Boucheron[1], and R. T. J. McAteer[2]

[1] Klipsch School of Electrical & Computer Engineering, New Mexico State University, Las Cruces, NM 88003, USA
e-mail: [amani;lboucher]@nmsu.edu
[2] Department of Astronomy, New Mexico State University, Las Cruces, NM 88003, USA
e-mail: mcateer@nmsu.edu

Received 26 February 2015 / Accepted 13 May 2015

ABSTRACT

Aims. We study the photospheric magnetic field of ~2000 active regions over solar cycle 23 to search for parameters that may be indicative of energy build-up and its subsequent release as a solar flare in the corona.
Methods. We extract three sets of parameters: (1) snapshots in space and time: total flux, magnetic gradients, and neutral lines; (2) evolution in time: flux evolution; and (3) structures at multiple size scales: wavelet analysis. This work combines standard pattern recognition and classification techniques via a relevance vector machine to determine (i.e., classify) whether a region is expected to flare (≥C1.0 according to GOES). We consider classification performance using all 38 extracted features and several feature subsets. Classification performance is quantified using both the true positive rate (the proportion of flares correctly predicted) and the true negative rate (the proportion of non-flares correctly classified). Additionally, we compute the true skill score which provides an equal weighting to true positive rate and true negative rate and the Heidke skill score to allow comparison to other flare forecasting work.
Results. We obtain a true skill score of ~0.5 for any predictive time window in the range 2 to 24 h, with a true positive rate of ~0.8 and a true negative rate of ~0.7. These values do not appear to depend on the predictive time window, although the Heidke skill score (<0.5) does. Features relating to snapshots of the distribution of magnetic gradients show the best predictive ability over all predictive time windows. Other gradient-related features and the instantaneous power at various wavelet scales also feature in the top five (of 38) ranked features in predictive power. It has always been clear that while the photospheric magnetic field governs the coronal non-potentiality (and hence likelihood of producing a solar flare), photospheric magnetic field information alone is not sufficient to determine this in a unique manner. Furthermore we are only measuring proxies of the magnetic energy build up. We are still lacking observational details on why energy is released at any particular point in time. We may have discovered the natural limit of the accuracy of flare predictions from these large scale studies.

Key words. methods: data analysis – techniques: image processing – Sun: flares – Sun: photosphere

1. Introduction

Solar flares are the result of magnetic energy release in the corona. As such, we require coronal magnetic field measurements in order to fully understand how this energy is built up (over days) and released (over minutes to hours). However, as such measurements are currently unavailable, much research has focused on inferring the coronal magnetic field structure from those data that are available, namely the photospheric magnetic field. It is assumed that turbulent motions on the surface of the Sun (the photosphere) twist and wind up the magnetic field in the corona. A complex photosphere causes a complex corona, a complex corona stores energy, and this stored energy is released as solar flares (McAteer et al. 2010). There have been many searches for a connection between coronal activity (flares) and the photospheric magnetic field (Ahmed et al. 2013; Falconer et al. 2011; Mason & Hoeksema 2010; Yuan et al. 2010; Huang et al. 2010; Jing et al. 2010, 2006; Yu et al. 2010a,b; Welsch et al. 2009; Ireland et al. 2008; Conlon et al. 2008; Georgoulis & Rust 2007; Leka & Barnes 2007; Schrijver 2007; Wang 2006; Barnes & Leka 2006; Guo et al. 2006; McAteer et al. 2005a; Abramenko 2005; Meunier 2004; Abramenko et al. 2003; Leka & Barnes 2003b; Hagyard et al. 1999; Zhang et al. 1994). Despite such research, a fully consistent and causal relationship has not yet been found (e.g., see conclusions in Mason & Hoeksema 2010; Leka & Barnes 2007; and Hagyard et al. 1999). Nevertheless, there is optimism that photospheric measures may yield some insight into imminent eruptive behavior (e.g., Falconer et al. 2011; Yu et al. 2010a; Schrijver 2007; Guo et al. 2006; Jing et al. 2006; Abramenko 2005; Leka & Barnes 2003b).

Many of the studies of the photospheric magnetic field extract a single parameter, or a few (≤7) parameters, and look for a relation to solar flare activity (Falconer et al. 2011; Mason & Hoeksema 2010; Yuan et al. 2010; Huang et al. 2010; Jing et al. 2010; Yu et al. 2010a; Ireland et al. 2008; Conlon et al. 2008; Georgoulis & Rust 2007; Schrijver 2007; Wang 2006; Guo et al. 2006; Jing et al. 2006; McAteer et al. 2005a; Abramenko 2005; Meunier 2004; Abramenko et al. 2003; Hagyard et al. 1999; Zhang et al. 1994). By using full vector magnetograms (Jing et al. 2010; Leka & Barnes 2007, 2003b; Barnes & Leka 2006; Hagyard et al. 1999; Zhang et al. 1994) to probe the transverse field it is possible to extract a larger number of parameters of magnetic complexity, albeit at the consequence of smaller datasets. Many of these studies focus on a predictive time window of 24 h (Ahmed et al. 2013; Falconer et al. 2011; Yuan et al. 2010; Jing et al. 2010; Leka & Barnes 2007; Schrijver 2007; McAteer et al. 2005a) although it is not clear that any of the extracted parameters are optimum for a 24-h predictive

time window. To the best of our knowledge, there are no studies that combine a large dataset with pattern recognition and classification techniques to study the time windows of predictions for each parameter.

In this paper, we will analyze line-of-sight (LOS) Michelson Doppler Imager (MDI) magnetograms (Scherrer et al. 1995) over solar cycle 23. We describe the complexity of each active region (AR) by extracting a large set of features of postulated importance for measuring the magnetic energy that has built up and compare these to the onset of solar flares over a range of time periods following each magnetogram. We explicitly include control data (i.e., regions that do not flare) in contrast to many studies of solely flaring regions (Schrijver 2007; Wang 2006; Guo et al. 2006; Meunier 2004; Abramenko et al. 2003; Hagyard et al. 1999; Zhang et al. 1994). We combine pattern recognition and classification techniques to determine (classify) whether a region is expected to flare; this is in contrast to many previous studies that rely merely on correlations or observations of parameters in relation to flaring activity (Jing et al. 2010, 2006; Ireland et al. 2008; Conlon et al. 2008; Georgoulis & Rust 2007; Schrijver 2007; Wang 2006; Guo et al. 2006; McAteer et al. 2005a; Abramenko 2005; Meunier 2004; Abramenko et al. 2003; Hagyard et al. 1999; Zhang et al. 1994).

Of particular interest is the characterization of local structure of AR fields. We consider three general categories of features. (1) Snapshots in space and time associated with increased flaring activity: (1a) total flux; (1b) magnetic gradients; (1c) neutral lines. (2) Evolution in time: flux evolution, which can act as energy release triggers. (3) Structures at multiple size scales: wavelet analysis, which allows separation of the field into its component lengthscales. Furthermore, we will consider these features in an automated classification framework whereby we will predict flare occurrence for a series of given time windows using a combination of all above-mentioned features.

The remainder of this paper is organized as follows. We discuss related work in the quantification of AR complexity in Sect. 2. We present details of the complexity features we use in Sect. 3, including a physical motivation for each of the features and the specifics of extracting those features. We briefly review automated classification methods and metrics for quantifying accuracy in Sect. 4, and present results using the proposed features for classification of flare activity and discuss discriminatory features in Sect. 5. Finally, we provide conclusions and future work in Sect. 6.

2. Analysis of active region complexity

In this section, we provide a detailed overview of related work on the use of AR complexity measures for prediction and characterization of solar flares. We focus on a listing of the specific features used in these studies (many of which we also use); accuracies, magnitudes of flares, and time windows (if published); and any significant conclusions regarding the use of photospheric complexity measures for flare prediction. We note here that it can be difficult to compare results from different papers as there are many metrics to quantify accuracy of flare prediction (several of which are defined in Sect. 4.3). Additionally, there are differing definitions of soft X-ray flux levels which constitute flaring versus non-flaring behavior and a range of predictive time windows considered.

The location of, and the gradients along, magnetic neutral lines play a key role in many studies. Ahmed et al. (2013) use machine learning to show that extensive properties connected to the neutral lines are closely connected to solar flares (≥C1.0) in a 24- or 48-h predictive time window, demonstrating flaring and non-flaring accuracies of 0.46 and 0.99, respectively. Falconer et al. (2011) uses the weighted length of the strong gradient neutral line, the magnetic area, and the length of the strong-field neutral line as proxies of free magnetic energy for prediction of flares (≥M1.0), coronal mass ejections, and high energy particle events over a 24-h predictive time window. Mason & Hoeksema (2010) use the total unsigned magnetic flux, primary inversion line (PIL) length, effective separation between the two polarities across the PIL, and the gradient-weighted inversion-line length (GWILL); they focus on the GWILL as the most promising measure for prediction of flares (≥M1.0) for a 6-h predictive time window, but find that it is "not a reliable parameter". Yuan et al. (2010), extending upon previous work in Song et al. (2009), use the total unsigned magnetic flux, length of the strong-gradient neutral line, and the total magnetic energy dissipation for flare forecasting (≥C1.0); they find weighted accuracies ranging from 0.65 to 0.86 in 24-h forecasts of flaring. Huang et al. (2010) use maximum horizontal gradient, length of the neutral line, and the number of singular points, incorporating temporal characteristics with the use of sequential supervised learning and voting by multiple classifiers; they achieve Heidke skill scores (HSS; see Sect. 4.3.2 for a definition) of approximately 0.65 for predictive flare index ≥M1.0 and for a 48-h predictive time window. Welsch et al. (2009) use many properties extracted from the magnetic field and flow field to associate the properties with flaring (≥C1.0) over 6- and 24-h windows via correlation and discriminant analysis; they find the unsigned flux near strong-field polarity inversion lines to be most strongly related to flare flux, yielding climatological skill scores ≤0.37. Song et al. (2009) use length of the strong gradient neutral line, total magnetic energy dissipation, and total unsigned magentic flux to forecast flares (≥C1.0, ≥M1.0, and ≥X1.0) within a 24-h time window; they find probabilities of detection of (0.90, 0.65, and 0.71) and false alarm rates of (0.29, 0.08, and 0.17), respectively. Schrijver (2007) proposes the use of the total unsigned flux "R" near high-gradient, strong-field polarity-inversion lines to characterize the electric currents in the photosphere; this parameter is found to have an increased value within a 24-h time window for large-flare (≥M1.0) producing ARs. Wang (2006) analyzes the relative motions of the two polarities of bipolar ARs and finds sudden change in magnetic shear following flares (≥M7.9). Guo et al. (2006) analyzes the effective distance (separation between flux-weighted centers of bipolar ARs), total flux, and tilt angle as compared to the Mount Wilson magnetic classification; they find that effective distance is well correlated with the Mount Wilson classes and with flaring activity (≥C1.0) in δ regions. Jing et al. (2006) use the mean of the spatial gradients along strong-gradient neutral lines, the length of the strong-gradient neutral lines, and the total magnetic energy dissipated in a unit layer and unit time and find positive correlation with flare (≥B1.0) activity.

Studies of local complexity across the AR have also shown some relation to solar flare activity. Yu et al. (2010b) use a wavelet transform to extract multiresolution features and use the same classifier as Huang et al. (2010), yielding an HSS of 0.77 for ARs with flare indices exceeding M1.0 and for a 48-h predictive time window. Yu et al. (2010a) use the same parameters and extract "sequential features" to characterize the temporal shapes of the features; a Bayesian network achieves HSS of 0.69, again for ARs with flare indices ≥M1.0 and for a 48-h predictive time window. Ireland et al. (2008) use statistics of the gradient distributions along multiscale opposite polarity region

separators and a Kolmogorov-Smirnov test to show that flaring (≥A1.0, ≥M1.0) and non-flaring regions come from different gradient distributions when considered over a 6-h time window. Conlon et al. (2008) use two measures of multifractality (contributional and dimensional diversity) along with total field strength and area to postulate a relationship between multifractal properties and flaring rate. Georgoulis & Rust (2007) define an effective magnetic field based on connectivity and show this provides a lower limit required for M-class flares and above. McAteer et al. (2005a) use the fractal dimension as determined with a modified box-counting algorithm and find that a large fractal dimension is a necessary but not sufficient condition for occurrence of large flares (≥C1.0) over a 24-h time window. Abramenko (2005) use structure functions to analyze multifractal properties of ARs and find that flaring regions tend to have larger degree of multifractality than do non-flaring regions.

Some authors have also considered features of the photospheric magnetic field extracted from vector magnetograms. Leka & Barnes (2003a), Leka & Barnes (2003b), Barnes & Leka (2006), and Leka & Barnes (2007) develop a comprehensive list of features derived from vector magnetograms, including measures from the distribution of magnetic fields, inclination angle, spatial gradient, vertical current density, twist, current helicity, shear angles, photospheric excess magnetic energy density, and magnetic charge topology models. They find mixed results, with some potential indicators of flare activity (≥M1.0) in Leka & Barnes (2003a), determining combinations of variables that indicate the ability to distinguish between flaring (≥M1.0) and non-flaring populations in Leka & Barnes (2003b), finding that coronal topology measures have better probabilities in distinguishing between flaring (≥C1.0) and non-flaring regions in Barnes & Leka (2006), and concluding that features based on the photospheric field have "limited bearing on whether that region will be flare productive" for flares ≥C1.0 and a 24-h predictive time window in Leka & Barnes (2007).

3. Image analysis

In this section we describe the three general categories of features used in this work: (1) snapshots in space and time, encompassing total magnetic flux, magnetic gradients, and neutral lines; (2) evolution in time, encompassing flux evolution; and (3) structures at multiple size scales, encompassing wavelet analysis. For each feature category, we first discuss the theoretical background and motivation followed by discussion of the image processing methods; we break up the discussion as such to better relate the image processing and theory.

For this work, we use MDI magnetograms from solar cycle 23, including NOAA ARs 8809–10 933, and some 260 000 total AR cutout images. ARs are selected based on locations specified by the Space Weather Prediction Center[1]. A 300″ × 300″ window is extracted centered on these locations. Data are cosine-corrected for line-of-sight effects, deprojected to a cylindrical equal-area mapping, and cropped to 211.5 Mm × 211.5 Mm (for details of these processes, see McAteer et al. 2005b). Additionally, magnetograms are considered only if the center of the AR is within 650″ of disk center to mitigate projection effects and disk edge artifacts; this leaves a total of 122 060 AR cutout images. It should be noted, however, that we have not implemented any correction for the saturation effect inherent in MDI data.

[1] www.swpc.noaa.gov/ftpmenu/forecasts/warehouse.html

Table 1. Extracted features.

Gradient features	**FE features**
Gradient mean	FE sum
Gradient std	FE absolute sum
Gradient median	FE gradient sum
Gradient min	FE 3σ area
Gradient max	FE mean
Gradient skewness	FE std
Gradient kurtosis	FE median
	FE min
Neutral line features	FE max
NL length	
NL no. fragments	**Wavelet features**
NL gradient-weighted length	Wavelet energy level 1
NL curvature mean	Wavelet energy level 2
NL curvature std	Wavelet energy level 3
NL curvature median	Wavelet energy level 4
NL curvature min	Wavelet energy level 5
NL curvature max	
NL bending energy mean	**Flux features**
NL bending energy std	Total unsigned flux
NL bending energy median	Total signed flux
NL bending energy min	Total negative flux
NL bending energy max	Total positive flux

3.1. Snapshots in space and time: total glux, gradient, and neutral line analysis

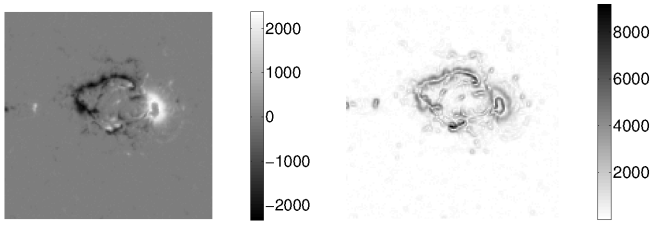
3.1.1. Theoretical background

At the photosphere the magnetic field is frozen into the plasma and advected by bulk plasma motions. Parker (1963) showed that energy can be stored in the corona when sunspots of opposite polarity are pushed together, forming an extended current sheet above the neutral line (NL). A shear flow has a similar effect in forming a current sheet above a NL. In both cases the NL often steadily lengthens until disrupted by some instability. As such, large magnetic gradients occur across the neutral line of large spots, particularly in the vicinity of large δ spots (Patty & Hagyard 1986; Zhang et al. 1994). Over a period of hours and days, the continued concentration of opposite polarities in a relatively small area leads to strong transverse gradients (Gallagher et al. 2002). We extract features related to the overall magnetic flux present; the gradient in magnetic flux across the AR; and the size and shape of and magnetic gradient along the NL.

3.1.2. Image processing

We compute a total of four features related to the magnetic flux as described here and summarized in Table 1. (1) The total unsigned magnetic flux is computed as the absolute sum of the magnetogram image. (2) The total signed magnetic flux is computed as the sum of the magnetogram image. (3) The total positive flux is computed as the sum of positive values of the magnetogram image. (4) The total negative flux is computed as the sum of negative values of the magnetogram image. These features describe the total magnetic flux present in the AR, as well as the flux imbalance in the region.

We compute a total of 7 features related to the gradient magnitude as described here. From the perspective of image processing, the spatial gradient is the first derivative of the image. This will highlight small specks and edges that may not be as visible in the original image. The gradient of image f at coordinates



(a) Magnetogram. (b) Gradient magnitude.

Fig. 1. Magnetogram and resultant gradient magnitude. NOAA AR # 10 488, 28 October 2003, 01:35. Colorbars are in units of Gauss and each image is 211.5 Mm × 211.5 Mm.

(x, y) is defined as the two-dimensional column vector (Gonzalez & Woods 2007)

$$v_f = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \mathrm{d}f/\mathrm{d}x \\ \mathrm{d}f/\mathrm{d}y \end{bmatrix} = \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)\hat{i} + \left(\frac{\mathrm{d}f}{\mathrm{d}y}\right)\hat{j}, \quad (1)$$

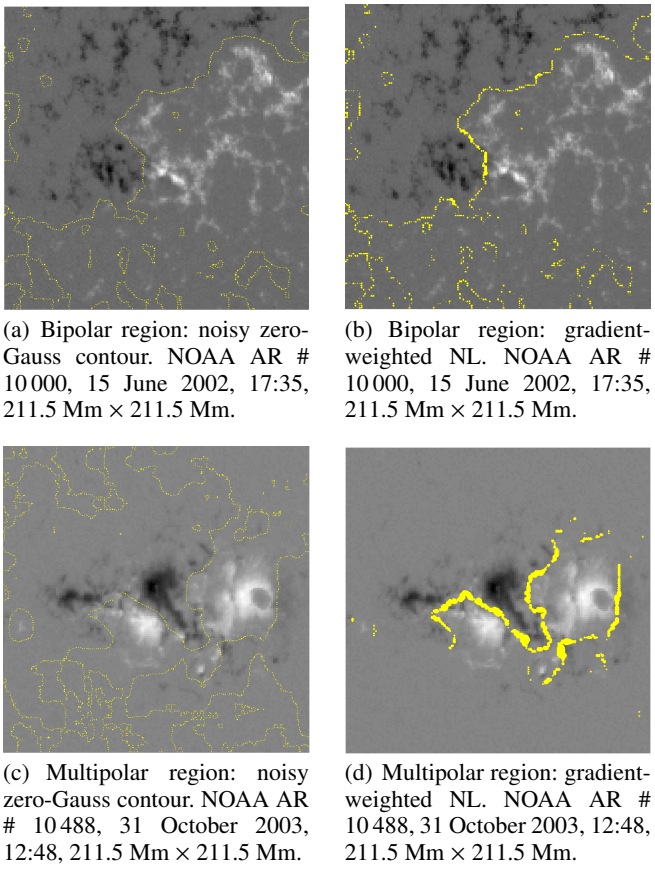
where G_x and G_y are the spatial gradients in the x and y directions, respectively, and $\hat{i}$ and $\hat{j}$ are unit vectors in the x and y directions, respectively. Gradient magnitude is defined as $|G(x, y)| = \sqrt{G_x^2 + G_y^2}$. To approximate the first derivative for discretely-indexed image f, we use the two Sobel filters

$$h_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad h_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix}. \quad (2)$$

The image is filtered (convolved) with each Sobel filter, yielding $G_x = h_x * f$ and $G_y = h_y * f$ where $*$ is the two-dimensional convolution operator. Figure 1 shows an example of one magnetogram image and the magnitude of the spatial gradient.

The gradient magnitude is computed for each magnetogram image in our dataset. To condense this gradient information into single descriptors (features) for each image, we compute the (1) mean; (2) standard deviation; (3) maximum; (4) minimum; (5) median; (6) skewness; and (7) kurtosis of the gradients in each image as summarized in Table 1. The gradient magnitude will be large for regions in which there are large differences in flux in close spatial proximity, and largest for opposite polarity regions with large flux in close proximity. These seven gradient features quantify the statistics of the occurrence of large gradient magnitude.

We compute a total of 13 features related to the NL as described here. The NL is detected in magnetogram images using the following procedure. First, magnetogram images are smoothed using a 10×10 pixel averaging filter to remove much of the statistical noise. Second, contours at the zero Gauss level of the smoothed image are used to create a NL mask. Third, since the zero-Gauss contours will flag all pixels at the zero-Gauss level, including those with very small gradient (i.e., including those of the quiet Sun), we mask the gradient image with the NL mask. This weights the NL to emphasize regions of the NL for which there is a large spatial gradient, indicating a transition between large positive and large negative flux. Figure 2 illustrates all of the zero-Gauss contours for both a bipolar and multipolar region, as well as a visualization of the gradient-weighted neutral line (GWNL) for the same bipolar and multipolar ARs. In this work, we make no distinction between the primary, high-gradient, or strong-field NL (Falconer et al. 2011; Yuan et al. 2010; Schrijver 2007; Jing et al. 2006) and the NL as it exists separating opposite polarity regions.



(a) Bipolar region: noisy zero-Gauss contour. NOAA AR # 10 000, 15 June 2002, 17:35, 211.5 Mm × 211.5 Mm.

(b) Bipolar region: gradient-weighted NL. NOAA AR # 10 000, 15 June 2002, 17:35, 211.5 Mm × 211.5 Mm.

(c) Multipolar region: noisy zero-Gauss contour. NOAA AR # 10 488, 31 October 2003, 12:48, 211.5 Mm × 211.5 Mm.

(d) Multipolar region: gradient-weighted NL. NOAA AR # 10 488, 31 October 2003, 12:48, 211.5 Mm × 211.5 Mm.

Fig. 2. Illustrations of NL analysis. **a)**, **c)** Noisy zero-Gauss contour before being weighted by the gradient image. We note the presence of zero-Gauss contours throughout the image, not just in the region separating the strong positive and negative flux. **b)**, **d)** Larger yellow markers indicate the presence of a stronger gradient at that spatial location along the NL.

The NL is detected for each image in our dataset. We define a total of 13 features related to the NLs. The first three features are defined as follows: (1) Length of the NL: the GWNL is thresholded at 20% of the maximum value to yield a strong-gradient binary NL mask. The NL length is determined as the sum of the pixels in this strong-gradient binary NL mask. (2) Number of fragments of the NL: using the same thresholded GWNL, the number of fragments is defined as the number of 8-connected components in the binary NL mask image. (3) GWNL length: the gradient-weighted length of the NL, computed by summing the pixels in the GWNL image. We note here that we define the strong-gradient NL in a relative image-by-image manner, as opposed to defining an absolute threshold for use across the dataset. We choose a relative threshold here as we feel it is important to define "strong" *relative* to the image in question; we will discuss this further in Sect. 5. These three features quantify different aspects of the NL length, which indicates the length of the region of most likely flux reconnection.

Additionally, ten more features are extracted based on the NL boundary curvature and NL bending energy. These features quantify the tortuousity of the NL boundary with the conjecture that a more tortuous NL indicates irregular and non-bipolar magnetic characteristics which may create more regions of probable flux reconnection. For extraction of these features, we trace the (closed) boundary of the NL in the thresholded GWNL image

and compute the orientation angle of each NL boundary pixel (Rodenacker & Bengtsson 2003):

$$\theta_n = \arctan\left(\frac{y(n+1) - y(n)}{x(n+1) - x(n)}\right), \; n = 1, \ldots, N \tag{3}$$

where x and y are the x- and y-coordinates of the N NL boundary pixels, and by definition $x(N+1) = x(1)$ and $y(N+1) = y(1)$. The curvature angles are computed separately for each NL segment, and the mean, standard deviation, maximum, minimum, and median are computed for all curvature angles for all NL segments; we forego the computation of skewness and kurtosis since there are too few data points for accurate computation of these higher-order moments. The bending energy B_e is analogous to the physical energy required to bend a rod and is computed as the normalized sum of the squared difference in curvature between subsequent boundary points (Rodenacker & Bengtsson 2003):

$$B_e = \frac{1}{N}\sum_{n=1}^{N}(\theta_{n+1} - \theta_n)^2 \tag{4}$$

where $\theta_{N+1} = \theta_1$ by definition. We note that the term "energy" in bending energy is distinct from the energy required to resist magnetic force in the AR. We use the bending energy as a measure of the shape of the NL and as a proxy for magnetic energy built up in the AR, motivated by the fact that NLs often underlie filaments which are often the site of coronal mass ejections associated with large flares. This measure is computed separately for each NL and the mean, standard deviation, maximum, minimum, and median are computed for the distribution of bending energy. Table 1 summarizes the 13 features extracted from NL analysis.

3.2. Evolution in time: flux evolution analysis

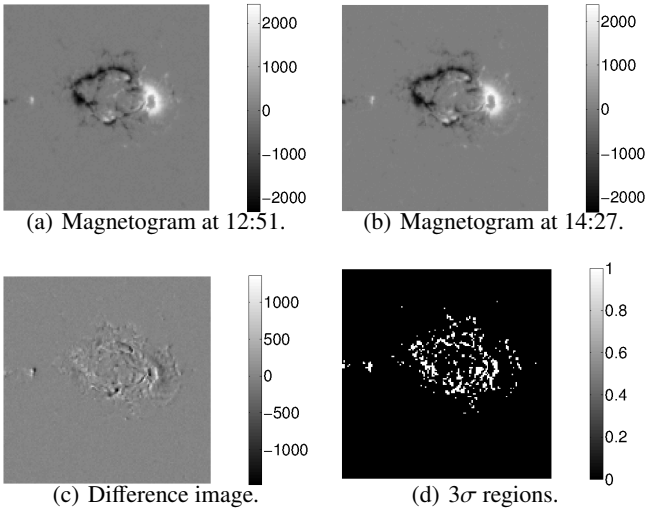
3.2.1. Theoretical background

Magnetic flux emergence is the origin of sunspots and ARs, and often is associated with solar eruptive events (Conlon et al. 2010). In the initial phase of AR emergence, the two opposite magnetic polarities move apart at a relatively large speed ($\sim$5 km s^{-1}) and then slow. New flux emerges continuously in the central part between the main polarities, separates and reaches the main polarities with high velocities. Emerging flux regions (EFRs) have been shown to have significance for solar flares (Tang & Wang 1993) and coronal mass ejections (Feynman & Martin 1995; Green et al. 2003). We extract features related to flux evolution in general, and a measure of emerging flux regions.

3.2.2. Image processing

Flux evolution is detected by considering difference images between two subsequent magnetograms, i.e., every two subsequent images (in one AR) are aligned and the previous magnetogram is subtracted from the following magnetogram to yield a difference image. To mitigate the effects of noise and to quantify strong changes, we identify regions in the difference image showing large deviations ($>3\sigma$) *above* the mean difference level. Again, we choose a relative (image-by-image) threshold at the 3σ level rather than an absolute threshold to better quantify large deviations on a per-image basis; this issue will be further discussed in Sect. 5. Figure 3 shows two subsequent images and the difference image along with a binary mask of the 3σ regions.

Nine features related to the difference image and flux evolution (FE) are extracted as follows and as summarized in Table 1.



(a) Magnetogram at 12:51. (b) Magnetogram at 14:27. (c) Difference image. (d) 3σ regions.

Fig. 3. Illustration of FE analysis. **a)**, **b)** Two magnetogram images, NOAA AR # 10 488, 29 October 2003, 12:51 and 14:27, 211.5 Mm × 211.5 Mm, **c)** the resultant difference image, **d)** binary mask of 3σ regions. Colorbars for **a)**–**c)** are in units of Gauss, and for **d)** is unitless.

(1) FE sum: the sum of the difference image (the FE image); (2) FE absolute sum: the sum of the absolute value of the FE image; (3) FE gradient: the sum of the gradient image of the second magnetogram masked by the binary 3σ image; (4) FE area 3 sigma: area of the 3σ regions. (5)–(9) FE mean, FE standard deviation, FE median, FE minimum, and FE maximum features: statistics of the FE image. It should be noted that features (1), (2), and (5)–(9) may have some contribution due to both flux emergence and submergence and features (3) and (4) explicitly characterize emerging flux regions.

3.3. Structures at multiple size scales: wavelet analysis

3.3.1. Theoretical background

A wavelet analysis of magnetograms discriminates spatially localized scale features such as the emergence/submergence of flux tubes. The wavelet transform maps scale content–the power in a particular location (Hewett et al. 2008). This is essential for determining the relative influence of local magnetic features against the global properties of the AR field. These localized magnetic features are fundamental to many flare theories and are important in developing our understanding of AR physics (Ireland et al. 2008). We extract features to quantify the structure of magnetic flux at different size scales by considering the high frequency edge content in different size scales. Large high-frequency edge content at a particular size scale indicates the presence of flux structures at a similar size scale. A large amount of smaller scale flux features could indicate a more complex magnetic structure with more chance for magnetic reconnection.

3.3.2. Image processing

The wavelet transform utilizes basis functions with compact time (spatial) support (i.e., finite in time/space). This is in contrast to the commonly used Fourier transform whose basis functions, complex exponentials, are not compact in time (space). Thus, the wavelet transform allows for both time (space) and frequency (wavenumber) resolution, although there is a tradeoff

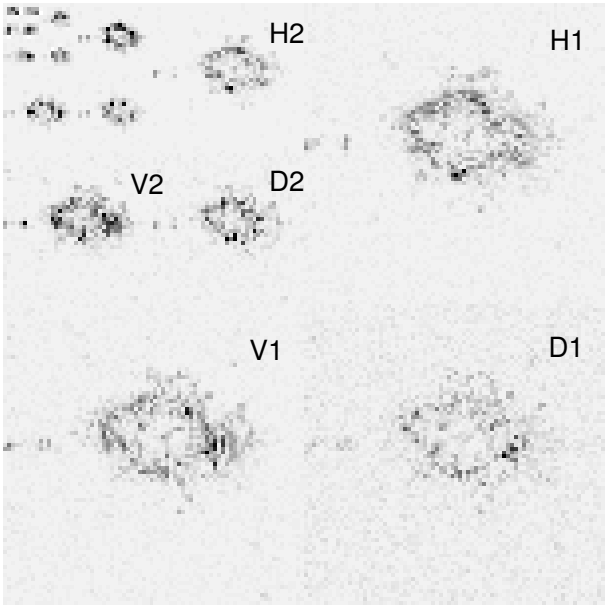


Fig. 4. Five-level wavelet decomposition for NOAA AR # 10 488, 28 October 2003, 01:35. The fifth-level decomposition is visualized in the uppermost left corner, with subsequently lower levels to the bottom right; as reference, the horizontal (H), vertical (V), and diagonal (D) images are labeled for the first and second level with the same structure in the subsequent levels. The pixel values (wavelet coefficients) can be considered unitless, with darker pixels representing higher values, scaled across all levels. The entire width of the image represents 211.5 Mm × 211.5 Mm; each subsequent decomposition reduces each dimension by half.

between the resolutions achievable simultaneously. A variety of different basis functions can be defined, each of which has different properties in time (space) and frequency (wavenumber). In this work, we use the Haar wavelet (Gonzalez et al. 2009) and 5 levels of decomposition.

Using the Haar transform, we can determine the resultant wavelet coefficients, yielding a low resolution image (which has been lowpass filtered and downsampled) and three highpass detail images (horizontal, vertical, and diagonal) for each level of decomposition. Each level of decomposition involves a downsampling operation in which each of the lowpass and highpass images are reduced in resolution by a factor of 2 in each dimension. Subsequent levels of decomposition begin with the lowpass image from the previous level. Figure 4 shows the five-level decomposition for an example magnetogram image, including the lowpass and three highpass detail images for each level. The lowpass image is a decimated (lowpass filtered and downsampled) version of the magnetogram image. The three highpass detail images are a highpass filtered and downsampled version of the magnetogram image. Since highpass filters will enhance edge structure in images, the highpass detail images contain information about the edge structure at the current image resolution, indicating the presence of magnetic flux elements at a similar resolution. These three highpass detail images are used to determine the energies of each decomposition level by summing the absolute values of the wavelet coefficients (the highpass images). We sum the energies of the three highpass images together as we are interested in an orientation independent measure of edge structure. We thus extract five energy values corresponding to each of the five levels of decomposition as summarized in Table 1.

4. Classification

In this section, we provide a brief overview of the classification method used in this work, our experimental setup, and the metrics with which we will assess performance. In the general formulation of classification, we wish to predict some discrete target variable t given some D-dimensional input vector $\boldsymbol{x}$ (Bishop 2006). In the work described here, target variable t corresponds to a decision that input data $\boldsymbol{x}$ belongs to flaring class C_1 or non-flaring class C_0. The equation $y(\boldsymbol{x}, \boldsymbol{w})$ which maps $\boldsymbol{x}$ to t is determined through optimization of some criterion based on training data.

4.1. Relevance vector machines (RVMs)

The Relevance Vector Machine (RVM; Tipping 2001, 2004; Tipping & Faul 2003) is a Bayesian sparse kernel technique for regression and classification which is a probabilistic generalization of the commonly used support vector machine (SVM; Burges 1998; Felzenszwalb et al. 2010; Melgani & Bruzzone 2004; Cao & Tay 2003; Tong & Koller 2002; Hua & Sun 2001; Furey et al. 2000; Chapelle et al. 1999; Drucker et al. 1999). In this formulation, classification is based on the function

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) \tag{5}$$

where y is a function whose sign indicates the class for a given D-dimensional input vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_D]^\top$; $\boldsymbol{w} = [w_0, w_1, \ldots, w_{M-1}]^\top$ is a vector of weights applied to the basis functions ϕ_j; basis functions ϕ_j are some linear or non-linear function of input data $\boldsymbol{x}$; and $\boldsymbol{\phi} = [\phi_0, \ldots, \phi_{M-1}]^\top$ (Bishop 2006). In this work, the 38-dimensional input vector $\boldsymbol{x}$ corresponds to the 38 features extracted for each AR image, The weight parameters $\boldsymbol{w}$ are chosen to optimize some criterion (the type-2 maximum likelihood in the case of RVMs), and the class indicated by $y(\boldsymbol{x}, \boldsymbol{w})$ indicates the prediction of flaring or not flaring. The weight vector $\boldsymbol{w}$ defines a decision boundary, a hyperplane in the multi-dimensional space spanned by $\boldsymbol{\phi}(\boldsymbol{x})$; data on opposite sides of this hyperplane are defined to belong to different classes. Since the function $y(\boldsymbol{x}, \boldsymbol{w})$ is linearly related to $\boldsymbol{w}$, the transformation $\boldsymbol{\phi}(\boldsymbol{x})$ allows for a non-linear decision boundary. We use the transformation $\boldsymbol{\phi}$ implicitly defined by the radial basis (Gaussian) kernel function $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}') = \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2 / 2\sigma^2)$.

4.2. Experimental setup

As output from the image analyses discussed in Sect. 3, we have one feature matrix per AR. We concatenate feature matrices for all ARs yielding feature matrix $\boldsymbol{X} = [X_1 X_2 ... X_N]^\top$ where N is the total number of ARs. X_i is the feature matrix of the ith AR with dimensionality $38 \times n_i$, where n_i is the number of images for the i-th AR; n_i is on the order of 150 for a typical AR. The 38 features encompass the total flux, gradient, neutral line, flux evolution, and wavelet features as discussed in Sect. 3. Our classification is based on consideration of this feature matrix $\boldsymbol{X}$ one row at a time, corresponding to one AR image which is considered one data point for classification. Within this formulation, we consider all data points (AR images) to be independent.

In addition to the feature matrix, training of supervised classifiers requires a label vector. In this application, each element in the label vector indicates whether the AR represented by those features will flare in the next k hours ("1") or not ("0"). Since

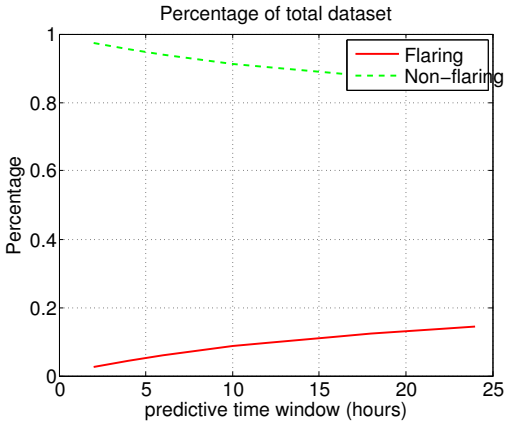


Fig. 5. Percentage of total dataset comprised of flaring regions and non-flaring regions, illustrating the unbalanced nature of this dataset.

Table 2. Flare forecasting confusion matrix (contingency table).

	Forecasted	
Observed	Flare	No flare
Flare	TP	FN
No flare	FP	TN

the predictive time window is larger than the nominal cadence of the MDI data (96 min), a single flare event will be predicted (classified) independently by multiple data points (AR images). Flares are defined from the geostationary operational environmental satellite (GOES) for flare magnitude ≥C1.0.

We randomly choose 1000 ARs to train the classifier and use a randomly subsampled balanced dataset from those 1000 ARs to find weight vector $\boldsymbol{w}$. A remaining 1000 ARs (and *all* associated data points) are used in testing, where weight vector $\boldsymbol{w}$ is used to predict the class (flare or no-flare) for each data point. Since a large majority of regions will not flare within the time window considered, classification methods can naively optimize their criterion by classifying all regions as class C_0 (no flare). For example, for a 4-h predictive time window, 4.5% of the total dataset belongs to flaring regions and 95.5% to non-flaring regions; this indicates that an overall accuracy of 95.5% can be achieved simply by predicting that none of the regions will flare. As a point of reference, we plot the percentage of the dataset comprised of flaring and non-flaring regions over all time windows in Fig. 5.

One method to alleviate the issue of unbalanced datasets is to artificially balance the dataset by subsampling one or both classes to be evenly represented. By cross validation across many randomly balanced datasets, we can get a better idea of accuracy. In this work, we use a 10-fold cross-validation with 500 samples each for flaring and non-flaring populations. Thus, we randomly subsample 500 flaring data points and 500 non-flaring data points from the 1000 ARs chosen for training; this process is repeated 10 times. Each classifier is tested on test data that has not been subsampled, consisting of 1000 ARs and some 60 000+ data points, to yield average accuracies. This 10-fold cross-validation is repeated for different predictive time windows in the interval [2, 24] h before flaring in a step of 2 h.

4.3. Metrics

The metrics we consider in this work can be derived from the basic confusion matrix (contingency table) shown in Table 2. TP is the true positive (correct flare forecast), FN false negative (incorrect no-flare forecast), FP false positive (incorrect flare forecast), and TN true negative (correct no-flare forecast).

4.3.1. True positive rate (TPR) and true negative rate (TNR)

Since flares are relatively rare events, overall classification accuracy (percentage of correctly classified data (TP + TN)/(TP + FN + FP + TN)) can be misleading. As such, we present both the percentage of correctly classified flaring regions (the true positive rate or TPR, also known as the sensitivity)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

and the percentage of correctly classified non-flaring regions (the true negative rate or TNR, also known as the specificity)

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}. \tag{7}$$

For use in further discussions, we also include the definition of the false negative rate (FNR),

$$\mathrm{FNR} = 1 - \mathrm{TPR} = \frac{\mathrm{FN}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

and the false positive rate (FPR),

$$\mathrm{FPR} = 1 - \mathrm{TNR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}. \tag{9}$$

4.3.2. Heidke skill score (HSS) and true skill score (TSS)

The use of skill scores attempts to mitigate issues of reporting classification accuracies for unbalanced data by combining all four terms of the confusion matrix. The Heidke skill score (HSS) and the Hanssen & Kuipers discriminant known as the true skill score (TSS) are the most widely used in flare forecasting (Bloomfield et al. 2012). HSS is defined as:

$$\mathrm{HSS} = \frac{2[(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FN} \times \mathrm{FP})]}{(\mathrm{TP} + \mathrm{FN})(\mathrm{FN} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FP})(\mathrm{FP} + \mathrm{TN})} \tag{10}$$

and TSS is given by:

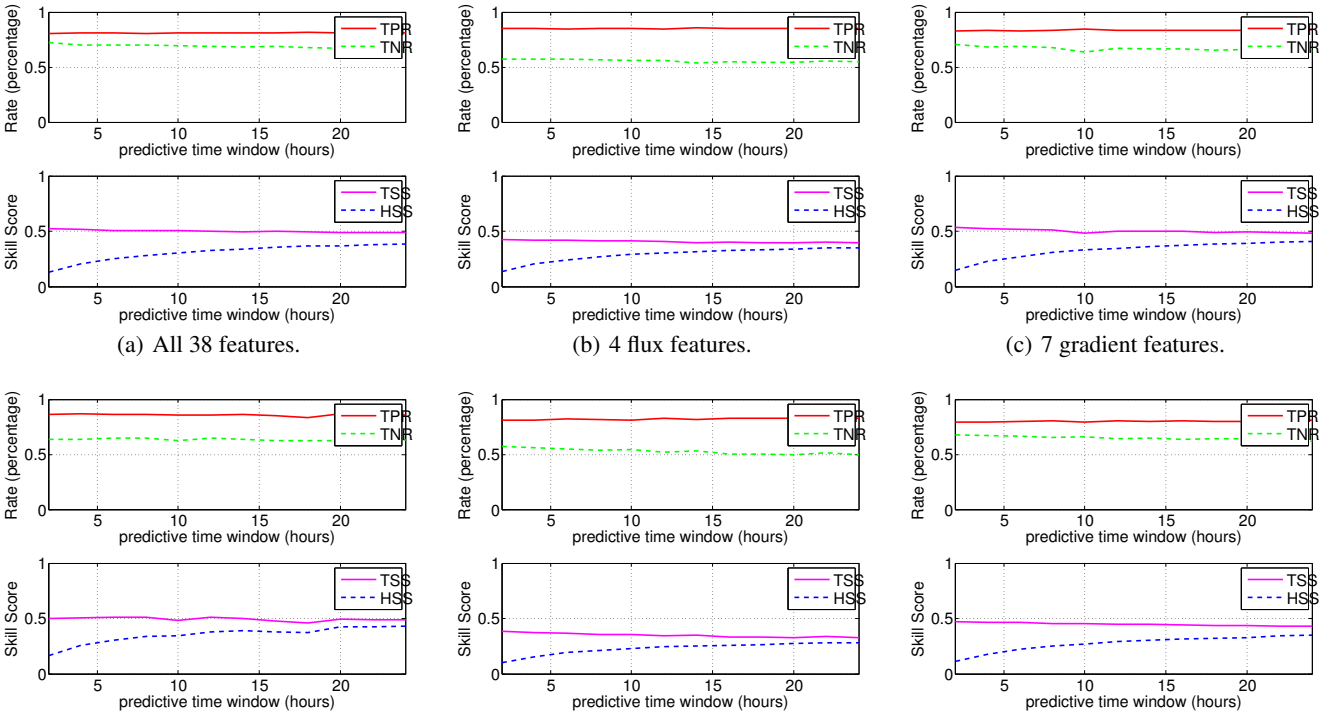
$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}. \tag{11}$$

We also note that TSS = TPR − FPR = TPR − (1 − TNR). Only TSS is unbiased for unbalanced datasets (Bloomfield et al. 2012).

5. Results

5.1. Classification using all features

Figure 6a shows the flaring (TPR) and non-flaring (TNR) accuracies and the skill score measures (HSS and TSS) for classification of 1000 randomly chosen ARs with respect to different predictive time windows using an RVM classifier and all 38 features. We see consistent performance across predictive time windows for both TPR (∼0.8) and TNR (∼0.7). We see consistent (∼0.5, perhaps slightly declining) TSS performance as the predictive time window increases, while HSS increases with predictive time window. We have also considered the standard deviation in performance across the 10 cross validation runs and



(a) All 38 features. (b) 4 flux features. (c) 7 gradient features.

(d) 13 neutral line features. (e) 9 flux evolution features. (f) 5 wavelet features.

Fig. 6. TPR (% correctly classified flaring regions), TNR (% correctly classified non-flaring regions), HSS, and TSS for RVM classification using different feature subsets.

found it to be 0.01–0.02 for TPR, 0.02–0.03 for TNR, 0.00–0.08 for HSS, and 0.01–0.02 for TSS. These small standard deviations indicate that the training set is of sufficient size for generalization of the trained RVM classifier to unseen test data. It is important to note that this classification considers a region as a flaring region if it flares *any* time within the predictive time window specified. Future work will consider a regression to determine *when* an AR is expected to flare; we expect that predictive time window will have a larger effect in regression analysis.

The increase in HSS with predictive time window is mainly due to its sensitivity to unbalanced datasets (as discussed in Bloomfield et al. 2012 and references therein). As predictive time window increases, the dataset becomes more balanced (see Fig. 5) while the underlying performance of the classifier (TPR and TNR) is largely unchanged. Since TPR and TNR are largely similar over the predictive time windows, the change in the four confusion matrix entries will have a much smaller effect than the change in dataset balance.

TSS may be slightly decreasing with increasing predictive time window, although it is not clear that this is a statistically significant trend. TSS may decrease if either TNR or TPR decrease; from Fig. 6, however, it appears that TNR is the most likely to be decreasing as predictive window increases. There are a variety of confounding factors which complicate the analysis of why TNR may be decreasing. As the predictive window increases, the dataset balance is changing, and entries from the non-flare row of the confusion matrix are moving to the flare row. If these entries were simply moving rows, we would expect TNR and TPR to move up or down in concert; we would additionally expect that TPR would move relatively more than TNR due to the imbalance of the dataset. We do not, however, observe this in Fig. 6. Thus, entries must be also moving between columns of the confusion matrix. This is not surprising since the RVM is now presented with different populations of data samples with which to optimize the decision boundary. In our case, it appears that entries are migrating in a fashion that has no noticeable effect on TPR, and a slight detrimental effect on TNR, indicating the non-flaring ARs are becoming more difficult to characterize as the predictive time window increases.

Table 3. Flare forecasting confusion matrix (contingency table) for all features, 4-h predictive time window.

	Forecasted	
Observed	Flare	No flare
Flare	TP = 2269	FN = 956
No flare	FP = 11 077	TN = 47 671

There are a variety of reasons that TNR could decrease, but we hypothesize that it is due to an ambiguity in predicting a non-flare. In predicting a flare, the features we use as a proxy of magnetic energy show a difference in regions that do flare versus those that do not. Once this change in features is noted, the region is more likely to flare. On the other hand, *lack* of a change in feature values at a specific point in time do not provide indication that these feature values will not change at a *future* point in time. This effect will be larger for larger predictive time windows. We note, however, that there is still a relatively minor degradation to either TNR or TSS over a large range of predictive time windows.

In order to determine the specific error (FN or FP) with the highest potential to improve the TSS performance, we show the confusion matrix for the 4-h predictive time window in Table 3. We note that the TPR term contributes positively to the TSS at a rate of 0.70 while the FPR contributes negatively by 0.19. While increasing either the TPR or decreasing the FPR (increasing the

Table 4. Top 5 features and TSS values (in parenthesis) for RVM classification using individual features.

Rank	2 h	4 h	6 h	8 h	10 h	12 h
1	G max (0.47)	G std (0.46) W L4 (0.46)	G std (0.47)	G std (0.46)	G std (0.45) W L4 (0.45	G std (0.45)
2	G std (0.46) W L3 (0.46) W L4 (0.46)	G max (0.45) W L3 (0.45) W L5 (0.45)	W L4 (0.46)	W L4 (0.45) W L3 (0.45)	W L3 (0.44) W L5 (0.44)	W L4 (0.44) W L3 (0.44)
3	W L5 (0.45) G mean (0.45) W L2 (0.45)	G mean (0.44) W L2 (0.44)	G max (0.45) W L3 (0.45)	G max (0.44) W L5 (0.44)	G mean (0.43) W L2 (0.43) G max (0.43	W L5 (0.43) G mean (0.43) W L2 (0.43)
4	FE std (0.34)	MF Neg (0.37)	W L2 (0.44) G mean (0.44) W L5 (0.44)	G mean (0.43) W L2 (0.43)	FE std (0.31)	G max (0.41)
5	MF neg (0.31)	FE std (0.32)	FE std (0.32)	MF Neg (0.34)	MF neg (0.29)	FE std (0.30)
Rank	14 h	16 h	18 h	20 h	22 h	24 h
1	G std (0.45)	G std (0.45)	G std (0.45)	G std (0.45)	G std (0.44)	G std (0.45)
2	W L3 (0.44) W L4 (0.44)	W L4 (0.43) W L3 (0.43) W L5 (0.43) W L2 (0.43) G mean (0.43)	W L4 (0.43) W L3 (0.43)	W L4 (0.43) W L3 (0.43)	W L4 (0.43) W L3 (0.43)	W L4 (0.43)
3	W L2 (0.43) W L5 (0.43)	G max (0.41)	W L5 (0.42) G mean (0.42) W L2 (0.42)	W L5 (0.42) W L2 (0.42) G mean (0.42)	W L5 (0.42) W L2 (0.42) G mean (0.42)	W L3 (0.42) W L5 (0.42) W L2 (0.42)
4	G mean (0.42)	FE std (0.29)	G max (0.41)	G max (0.41)	G max (0.39)	G mean (0.41)
5	G max (0.41)	MF neg (0.25)	FE std (0.29)	FE std (0.30)	FE std (0.29)	G max (0.39)

Notes. MF are magnetic flux features, G gradient features, NL neutral line features, FE flux evolution features, and W wavelet features.

TNR) can improve the TSS, there will be more advantage to improving the TPR since that is the variable with the most room for improvement. We note, however, that the consequence of each of the two errors FP and FN may be significantly different, which may justify more focus to improving either TPR or FPR, independent of the room for improvement. Of course, since all four measures in the confusion matrix are inherently related, it may be difficult to improve TPR without negatively affecting FPR. We will discuss several potential modifications to the classification process that may improve TSS in Sect. 6.

5.2. Classification using feature subsets

We consider the classification performance using subsets of features by training and testing an RVM on a subset of features. For example, we train and test using the 4 flux features and achieve performance as shown in Fig. 6b. In a similar manner, classification using other feature subsets are shown in Figs. 6c–f. Our goal here is to determine which subsets may have better accuracy and to allow for future work in postulating the physical relation between features and AR flaring. We note similar performance for TPR, between 0.80 and 0.85. There are significant differences, however, for the TNR performance of different feature subsets, ranging from 0.45 to 0.69. We find standard deviation in performance across the 10 cross validation runs to be 0.01–0.06 for TPR, 0.01–0.11 for TNR, 0.00-0.09 for HSS, and 0.00–0.07 for TSS. These ranges in standard deviation consider the range across all of the feature subsets. In general, the poorer performing feature subsets tend to display a larger standard deviation.

These differences in performance can be considered simultaneously in the TSS plots (since TSS is linearly related to both TPR and TNR). In particular, we note that the gradient features yield the highest TSS, while the FE features yield the lowest. Indeed, the gradient features alone yield performance very similar to that of all the features combined. We note a similar trend in the feature subset results to the 38-feature results in that performance is largely similar across the range of predictive time windows.

5.3. Classification using individual features

As a further study of the discriminatory potential of specific features, we consider the classification performance using single features as input to the RVM. As in the experiments with feature subsets, we train and test an RVM on a single feature. The performance over features and predictive time windows is summarized in Table 4 where we show the top five features ranked according to their TSS. We find much consistency in the top ranked features, with the standard deviation of the spatial gradient being the top-ranked feature for all but the 2-h predictive time window. Other commonly occurring features include the maximum gradient, mean gradient, and the various wavelet energies. We find no significant differences in discriminatory features across the different predictive time windows.

We make three observations regarding the performance of individual features. First, we note that all features with TSS > 0.40 are either gradient or wavelet features. This indicates that the most discriminatory features come from either gradient analysis or wavelet analysis which, at a basic level, quantify edge strengths in the magnetogram. Second, we note that the various statistics of the gradient image and the energies of the various size scales of the wavelet analysis provide largely the same discriminatory potential. Third, we note that it is interesting that the gradient standard deviation alone can achieve a TSS close to that of the classification using all features. It is important to note, however, that the individual feature performance does not indicate the discriminatory potential for a feature when *combined* with other features (as in the feature subset plots in Fig. 6). In future work, we will consider the use of optimal subsets of features, as further discussed in Sect. 6.

Table 5. Comparison to related flare prediction methods.

Reference	ARs	Images	Flares[a]	Magnitude	Window (h)	Temporal	TPR	TNR	TSS	HSS
Ours	2124	122 060	3432–19 086[b]	≥C1.0	2–24	No	0.81	0.70	0.51	0.39
1	N/A[c]	?[d]	8498	≥C1.0	24	No	0.46[e]	0.99[e]	0.45[f]	0.54[e]
2	230	?[d]	167	≥C1.0	24	No	0.33[g]	0.92[g]	0.25[g]	0.29[g]
3	870	48 344	8612	≥M1.0[h]	48	Yes	0.90[i]	0.88[i]	0.78[f]	0.66[i]
4	?[d]	31 164	8510	≥M1.0[h]	48	Yes	0.85[j]	0.88[j]	0.73[f]	0.69[j]
5	1010	55 582	9801	≥M1.0[h]	48	Yes	0.95[k]	0.92[k]	0.87[f]	0.77[k]
6	46	2708	119 228[b]	≥C1.0	6, 24	No	0.49[l]	0.99[l]	0.47[l]	0.51[l]
7	?[d]	55	54	≥C1.0	24	No	0.75[m]	0.93[m]	0.68[m]	0.69[m]

Notes. [(a)] Number of data points associated with a flare; multiple data points may include the same flare within the time window. [(b)] The range of values is due to the range in time windows. [(c)] Magnetic features are considered rather than ARs. [(d)] This information was not readily apparent from the paper. [(e)] From Table 5 in Ahmed et al. (2013). [(f)] Computed from TPR and TNR. [(g)] Compiled from Figs. 3–6 in Yuan et al. (2010). [(h)] This magnitude specifies the total flare importance index. [(i)] From Fig. 5 in Huang et al. (2010). [(j)] From Table 5, BN_F column in Yu et al. (2010b). [(k)] From Table 5, MODWT_DB2_Red column in Yu et al. (2010a). [(l)] Computed from Table 4, 24N FLCT in Welsch et al. (2009). [(m)] Computed from Table 8, Model (7) in Song et al. (2009).

References. (1) Ahmed et al. (2013); (2) Yuan et al. (2010); (3) Huang et al. (2010); (4) Yu et al. (2010b); (5) Yu et al. (2010a); (6) Welsch et al. (2009); (7) Song et al. (2009).

5.4. Relative versus absolute thresholds

In this work, we chose to use relative thresholds for segmenting the NL (20% of the maximum value of the gradient-weighted NL) and the FE 3σ regions (3σ above the mean value of the different image). To study the effect of using relative versus absolute thresholds for computation of these features, we ran the classification simulations with features computed using two different absolute thresholds. In the first case, we chose an absolute threshold for both the NL and FE features to be the mean of the relative threshold across the entire dataset, resulting in thresholds of 384 G for the NL and 54 G for the 3σ regions. In the second case, we used a threshold for both the NL and FE features of 50G, a common threshold used in the literature for "strong" flux.

In the first case (384 G for NL and 54 G for FE), we see a slight increase in performance for the NL features which also positively affected the results for all features. In particular, we note an increase in TSS of approximately 0.03 across the predictive time windows for both the NL features alone and for the 38-feature results. The use of this absolute threshold had no noticeable effect on the TSS of the FE features. In the second case (50 G for both NL and FE), we see a slight decrease in performance for the NL features of approximately 0.03 in TSS. The decreased performance of the NL features did not affect the overall 38-feature results. The use of this absolute threshold had no noticeable effect on the TSS of the FE features. It is unclear whether any of these differences in performance are statistically significant, as they are only slightly larger than the 0.02 standard deviation in performance measured across the 10 cross validation runs. Future work will consider in more detail the effects of relative versus absolute thresholds.

5.5. Comparison to related work

We now discuss our results in light of results published in related work, particularly those with quantitative metrics of performance (Ahmed et al. 2013; Yuan et al. 2010; Huang et al. 2010; Yu et al. 2010a,b; Welsch et al. 2009; Song et al. 2009). As mentioned in Sect. 2, use of different datasets, accuracy metrics, flare magnitudes, and time windows can complicate direct comparision of results. In this discussion, we highlight similarities and differences in the methods as well as datasets, metrics, magnitudes, and time windows. Additionally, we summarize some key characteristics of the dataset and performance in Table 5.

We discuss here some of the similarities and differences in the datasets for the aforementioned work. First, we note that our dataset, at 2124 ARs and 122 060 total images is over twice the size of the largest dataset considered in other work besides Ahmed et al. (2013), which considers magnetic features rather than NOAA ARs. Second, we note that the magnitude considered to constitute a flare is ≥C1.0 for our work, Ahmed et al. (2013), Yuan et al. (2010), Welsch et al. (2009), and Song et al. (2009); other work considers regions with a total flare *importance index* of ≥M1.0 (Huang et al. 2010; Yu et al. 2010a,b). Third, we consider a range of predictive time windows from 2 to 24 h; other work considers 6 h (Welsch et al. 2009), 24 h (Ahmed et al. 2013; Yuan et al. 2010; Welsch et al. 2009; Song et al. 2009), or 48 h (Ahmed et al. 2013; Huang et al. 2010; Yu et al. 2010a,b). Fourth, we note that some researchers have begun using temporal information for flare prediction (Huang et al. 2010; Yu et al. 2010a,b).

Compared to those works that do not use temporal information (Ahmed et al. 2013; Yuan et al. 2010; Welsch et al. 2009; Song et al. 2009), we find our method to have a higher TPR (0.81 versus 0.26–0.49), lower TNR (0.70 versus 0.96–0.99), higher TSS (0.51 versus 0.22–0.47), and higher HSS (0.39 versus 0.12–0.22). Exceptions to this trend are the method of Song et al. (2009), which was applied on a very small dataset of 65 samples, and the method of Ahmed et al. (2013) which has a higher HSS due to lack of dataset balancing. Ahmed et al. (2013), Yuan et al. (2010), and Welsch et al. (2009) do not appear to balance their datasets prior to classification, which likely skews their accuracies in favor of the majority negative class. Compared to those works wich do use temporal information (Huang et al. 2010; Yu et al. 2010a,b), we find our method to have lower TPR (0.81 versus 0.85–0.95) and lower TNR (0.70 versus 0.88–0.98). There are three potential sources for this difference in performance: the different flare magnitudes, the use of temporal features, and different size datasets. We will consider the implementation of temporal features as well as study the effect of different flare magnitudes in future work as we will discuss in Sect. 6.

6. Conclusion and future work

We used a large set of LOS magnetograms, including ARs which ultimately flared and control ARs which did not flare. We extracted 38 different features related to the complexity of each AR magnetogram. These features resulted from an analysis of the total flux, spatial gradient, NL, flux evolution, and wavelet decomposition. This is the largest scale study carried out to date, in terms of combining a large number of features and a large dataset. An RVM standard pattern recognition framework was used to classify whether the given AR will produce a solar flare. In general, we achieved TPRs of ~0.8 and TNRs of ~0.7. These rates correspond to a TSS of approximately 0.5. In comparison to other studies of flare prediction from static images (Ahmed et al. 2013; Yuan et al. 2010; Huang et al. 2010; Yu et al. 2010a,b; Welsch et al. 2009; and as summarized in Bloomfield et al. 2012), we find our classification performance to be higher for TNR and lower for TPR. However, in a comparison to studies that use temporal features, the TPR and TNR discovered here are slightly lower. The TNR and TPR do not vary much when predicting over the 2–24 h window.

Upon ranking, features related to magnetic gradients are associated with the best predictive ability. Features related to power at various wavelet scale decompositions also feature in the top 5. This agrees with and improves upon previous work, where the size scale of the neutral line was studied (Ireland et al. 2008).

The large size of this study, compared to previous work, only resulted in small improvements over previous work. This naturally leads us to question where future advances can be made. Clearly, just adding in more data and more features is not necessarily the best approach. It has always been clear that while the photospheric magnetic field governs the coronal non-potentiality (and hence likelihood to produce a solar flare), photospheric magnetic field information alone is not sufficient to determine coronal structure. Chromospheric, and eventually coronal, magnetic field is required. In addition, we emphasize that this type of study is only measuring a proxy of the magnetic energy *build up*. We are still lacking observational details on why energy is released at any particular point in time. It is also unclear, both observationally and theoretically as to how much (i.e., what fraction) of stored energy is released (McAteer et al. 2007; McAteer & Bloomfield 2013), and how this is distributed between thermal emission, non-thermal emission, and bulk motions (Emslie et al. 2012). With this in mind, we may have discovered the natural limit of the accuracy of flare predictions from these large scale photospheric studies.

However, some further advances can be made with current data. We plan for three further investigations related to the features themselves. First, we will implement fractal dimension features, to include computation of the fractal dimension of the NL and features related to the grayscale fractal dimension of the magnetogram itself. Fractality and multifractality has been shown to be a highly discriminative feature in other application domains (Spillman Jr. et al. 2004; McAteer 2013) and warrants further investigation (Conlon et al. 2008; McAteer 2015; McAteer et al. 2015). As a second investigation related to the complexity features, we will consider the use of feature selection methods. While we have considered arbitrary feature subsets according to the image processing methods (e.g., gradient or NL), automated methods can determine optimal (or close to optimal) feature subsets (Pudil et al. 1994). This analysis will allow insight into the specific physical processes that directly precede flares. As a third exploration, we will consider the implementation of temporal features to characterize the change in appearance of ARs leading up to a flare.

We also plan for four further investigations related to the pattern recognition aspects of our work. First, we will repeat this experiment for a variety of flare sizes (e.g., C1.0, C5.0, M1.0, M5.0, X1.0) to study and mitigate the bias associated with high solar backgrounds. Second, we will investigate other means of classifying our unbalanced dataset. This work used a cross-validation framework in which the dataset was subsampled to yield equal contribution from flaring and non-flaring regions. As the different errors (FP and FN) have very different implications in flare forecasting, we can implement a cost matrix which applies a different penalty to the two different errors. This may allow us to tune the performance to a better suited level for flare prediction. Third, we will implement these features in a regression analysis (using RVMs) where we will predict *when* a flare will occur rather than the binary decision that a flare will occur within some timeframe. This will provide further insight into the predictive time windows associated with flare prediction, and which features may be more applicable across the different predictive time windows. Fourth, we will analyze the classification and regression frameworks for prediction of other solar eruptive events often coincident with solar flares, including coronal mass ejections and solar energetic particles.

Acknowledgements. The authors gratefully acknowledge an NMSU Vice President for Research Interdisciplinary Research Grant, NSF PAARE grant AST-0849986, NASA EPSCoR grant NNX09AP76A, and NNH09CE72C, all of which helped support this work. One of us (JMA) was partially supported by a National Science Foundation Career award NSF AGS-1255024, and NASA contracts NNH12CG10C and NNX13AE03G.

References

Abramenko, V. I. 2005, Sol. Phys., 228, 29
Abramenko, V. I., Yurchyshyn, V. B., Wang, H., Spirock, T. J., & Goode, P. R. 2003, ApJ, 597, 1135
Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, Sol. Phys., 283, 157
Barnes, G., & Leka, K. D. 2006, ApJ, 646, 1303
Bishop, C. M. 2006, Pattern Recognition and Machine Learning (New York: Springer)
Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJ, 747, L41
Burges, C. J. C. 1998, Data Mining and Knowledge Discovery, 2, 121
Cao, L.-J., & Tay, F. E. H. 2003, IEEE Transactions on Neural Networks, 14, 1506
Chapelle, O., Haffner, P., & Vapnik, V. N. 1999, IEEE Transactions on Neural Networks, 10, 1055
Conlon, P. A., Gallagher, P. T., McAteer, R. T. J., et al. 2008, Sol. Phys., 248, 297
Conlon, P. A., Gallagher, P. T., McAteer, R. T. J., & Fennell, L. 2010, ApJ, 722, 577
Drucker, H., Wu, S., & Vapnik, V. N. 1999, IEEE Transactions on Neural Networks, 10, 1048
Emslie, A., Dennis, B., Shih, A. Y., et al. 2012, ApJ, 759, 71
Falconer, D., Barghouty, A. F., Khazanov, I., & Moore, R. 2011, Space Weather, 9, 04003
Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. 2010, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 1627
Feynman, J., & Martin, S. F. 1995, J. Geophys. Res., 100, 3355
Furey, T. S., Cristianini, N., Duffy, N., et al. 2000, Bioinformatics, 16, 906
Gallagher, P. T., Moon, Y.-J., & Wang, H. 2002, Sol. Phys., 209, 171
Georgoulis, M. K., & Rust, D. M. 2007, ApJ, 661, L109
Gonzalez, R. C., & Woods, R. E. 2007, Digital Image Processing (New Jersey: Prentice Hall)
Gonzalez, R. C., Woods, R. E., & Eddins, S. L. 2009, Digital Image Processing Using Matlab (Gatesmark Publishing)
Green, L. M., Démoulin, P., Mandrini, C. H., & Van Driel-Gesztelyi, L. 2003, Sol. Phys., 215, 307
Guo, J., Zhang, H., Chumak, O. V., & Liu, Y. 2006, Sol. Phys., 237, 25
Hagyard, M. J., Stark, B. A., & Venkatakrishnan, P. 1999, Sol. Phys., 184, 133

Hewett, R. J., Gallagher, P. T., McAteer, R., Ireland, J., & Young, C. 2008, Sol. Phys., 248, 297
Hua, S., & Sun, Z. 2001, Bioinformatics, 17, 721
Huang, X., Yu, D., Hu, Q., Wang, H., & Cui, Y. 2010, Sol. Phys., 263, 175
Ireland, J., Young, C. A., McAteer, R. T. J., et al. 2008, Sol. Phys., 252, 121
Jing, J., Song, H., Abramenko, V., Tin, C., & Wang, H. 2006, ApJ, 644, 1273
Jing, J., Tin, C., Yuan, Y., et al. 2010, ApJ, 713, 440
Leka, K. D., & Barnes, G. 2003a, ApJ, 595, 1277
Leka, K. D., & Barnes, G. 2003b, ApJ, 595, 1296
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Mason, J. P., & Hoeksema, J. T. 2010, ApJ, 723, 634
McAteer, R. T. J. 2013, in SOC and Fractal Geometry in Self Organized Criticality Systems, 1, 73
McAteer, R. T. J. 2015, Sol. Phys., in press
McAteer, R., & Bloomfield, D. 2013, ApJ, 776, 66
McAteer, R. T. J., Gallagher, P. T., & Ireland, J. 2005a, ApJ, 631, 628
McAteer, R. T. J., Gallagher, P. T., Ireland, J., & Young, C. A. 2005b, Sol. Phys., 228, 55
McAteer, R., Young, C., Ireland, J., & Gallagher, P. 2007, ApJ, 662, 691
McAteer, R. T. J., Gallagher, P. T., & Conlon, P. 2010, Adv. Space Sci. Res., 45, 1067
McAteer, R. T. J., Aschwanden, M., Dimitropoulou, M., et al. 2015, Space Sci. Rev., in press
Melgani, F., & Bruzzone, L. 2004, IEEE Transactions on Geoscience and Remote Sensing, 42, 1778
Meunier, N. 2004, A&A, 420, 333
Parker, E. N. 1963, ApJ, 138, 552
Patty, S. R., & Hagyard, M. J. 1986, Sol. Phys., 103, 111
Pudil, P., Novovičová, J., & Kittler, J. 1994, Pattern Recognition Letters, 15, 1119
Rodenacker, K., & Bengtsson, E. 2003, Analytical Cellular Pathology, 25, 1
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, Sol. Phys., 162, 129
Schrijver, C. J. 2007, ApJ, 655, L117
Song, H., Tan, C., Jing, J., et al. 2009, Sol. Phys., 254, 101
Spillman Jr, W. B., Robertson, J. L., Huckle, W. R., Govindan, B. S., & Meissner, K. E. 2004, Phys. Rev. E, 70, 061911
Tang, F., & Wang, H. 1993, Sol. Phys., 143, 107
Tipping, M. E. 2001, J. Machine Learning Res., 1, 211
Tipping, M. E. 2004, Bayesian inference: An introduction to principles and practice in machine learning (Berlin, Heidelberg: Springer), 41
Tipping, M. E., & Faul, A. C. 2003, in The Ninth International Workshop on Artificial Intelligence and Statistics, 1
Tong, S., & Koller, D. 2002, J. Machine Learning Res., 2, 45
Wang, H. 2006, ApJ, 649, 490
Welsch, B. T., Li, Y., Schuck, P. W., & Fisher, G. H. 2009, ApJ, 705, 821
Yu, D., Huang, X., Hu, Q., et al. 2010a, ApJ, 709, 321
Yu, D., Huang, X., Wang, H., et al. 2010b, ApJ, 710, 869
Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. 2010, RA&A, 10, 785
Zhang, H., Ai, G., Yan, X., Li, W., & Liu, Y. 1994, ApJ, 423, 828