# Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms

Xin Huang[1], Huaning Wang[1,2] , Long Xu[1], Jinfu Liu[3], Rong Li[4], and Xinghua Dai[1]

[1] Key Laboratory of Solar Activity, National Astronomical Observatories of Chinese Academy of Sciences Beijing, People's Republic of China; xhuang@bao.ac.cn
[2] School of Astronomy and Space Science University of Chinese Academy of Sciences Beijing, People's Republic of China
[3] Harbin Institute of Technology Harbin, People's Republic of China
[4] School of Information Beijing Wuzi University Beijing, People's Republic of China

## Abstract

Solar flares originate from the release of the energy stored in the magnetic field of solar active regions, the triggering mechanism for these flares, however, remains unknown. For this reason, the conventional solar flare forecast is essentially based on the statistic relationship between solar flares and measures extracted from observational data. In the current work, the deep learning method is applied to set up the solar flare forecasting model, in which forecasting patterns can be learned from line-of-sight magnetograms of solar active regions. In order to obtain a large amount of observational data to train the forecasting model and test its performance, a data set is created from line-of-sight magnetogarms of active regions observed by *SOHO*/MDI and *SDO*/HMI from 1996 April to 2015 October and corresponding soft X-ray solar flares observed by *GOES*. The testing results of the forecasting model indicate that (1) the forecasting patterns can be automatically reached with the MDI data and they can also be applied to the HMI data; furthermore, these forecasting patterns are robust to the noise in the observational data; (2) the performance of the deep learning forecasting model is not sensitive to the given forecasting periods (6, 12, 24, or 48 hr); (3) the performance of the proposed forecasting model is comparable to that of the state-of-the-art flare forecasting models, even if the duration of the total magnetograms continuously spans 19.5 years. Case analyses demonstrate that the deep learning based solar flare forecasting model pays attention to areas with the magnetic polarity-inversion line or the strong magnetic field in magnetograms of active regions.

*Key words:* methods: data analysis – Sun: activity – Sun: flares – techniques: image processing

*Supporting material:* animation

## 1. Introduction

Solar flares associated with energetic charged particles and electromagnetic radiation could affect the radio communication, the precision of Global Positioning Systems, and the safety of satellites and astronauts within several minutes. In order to reduce or prevent damages of solar flares, the development of solar flare forecasting technology has become very important for the modern society supported with high-tech systems.

Since the 1930s the statistical relationship between solar flares and measures has been widely employed in solar flare forecasting technology. Giovanelli (1939) suggested a solar flare forecasting model based on the statistical relationship between solar flares and sunspots. And then, more and more statistical methods were applied in solar flare forecasting models (Jakimiec 1987; Bornmann & Shaw 1994). The Poisson statistical model was developed to estimate the probability of solar flares in Gallagher et al. (2002) and Bloomfield et al. (2012). Leka & Barnes (2003) and Barnes et al. (2007) applied the discriminant analysis to determine the importance of magnetic parameters for solar flares. Mason & Hoeksema (2010) used the superposed epoch analysis method

to find a statistical relationship between magnetic field parameters and solar flares.

With the development of computer technology, artificial intelligence methods have also been used to build computer-aided solar flare forecasting models. At the early stage of artificial intelligence, the expert system WOLF (Miller 1988) was used to forecast the occurrence of solar flares. Actually, an expert system, which emulates the decision-making of a human expert, is limited by its knowledge acquisition ability. Therefore, the machine learning method is developed to discover the knowledge from data. From now on, many machine learning algorithms have played important roles in solar flare forecasts, such as the support vector machine (Li et al. 2007; Bobra & Couvidat 2015; Nishizuka et al. 2017), the artificial neural network (Qahwaji & Colak 2007; Wang et al. 2008; Colak & Qahwaji 2009; Ahmed et al. 2013), the ordinal logistic regression (Song et al. 2009), the Bayesian network approach (Yu et al. 2010), and the ensemble learning method (Guerra et al. 2015). These approaches were used to forecast whether or not a flare will happen. By using the regressor, continuous flare size can be forecasted in Boucheron et al. (2015) and Muranushi et al. (2015).

The above-mentioned statistical and machine learning solar flare forecasting models mainly rely on morphological or physical parameters extracted from active regions, and many researchers have tried their best to find effective forecasting measures. A classification of sunspots was proposed by McIntosh (1990), and Lee et al. (2012) explored the relationship between classifications of sunspots and solar flares. It is

obvious that the morphological classification of sunspots is artificial, and physical parameters are more reasonable measures for the complexity and nonpotentiality of active regions, for example, the length of the neutral line (Falconer 2001), the gradient of the magnetic field (Cui et al. 2006), the highly stressed longitudinal magnetic field (Huang & Wang 2013), the distance between active regions and predicted active longitudes (Huang et al. 2013), the Zernike moment of magnetograms (Raboonik et al. 2016). Furthermore, The evolutions of physical parameters in active regions were studied in Yu et al. (2009, 2009), Huang et al. (2010), and Korsós et al. (2014, 2015). Besides that, Wheatland (2004) suggested that the history of solar flares is an important predictor for subsequent flares. However, most of the measures extracted from active regions are strongly correlated with each other, and none of them significantly outperformed all of the others (Leka & Barnes 2007; Barnes & Leka 2008; Barnes et al. 2016). Up to now, the extraction of the effective forecasting measures from action regions has become a bottleneck to improve the performance of solar flare forecasting models.

All of the forecasting measures are extracted from two-dimensional solar observational data, and are partially related to solar flares. It is obvious that there exist some patterns related to solar flares in the observational data. The deep learning method (Hinton & Salakhutdinov 2006; Arel et al. 2010; LeCun et al. 2015; Schmidhuber 2015), which is one of the fastest-growing and most exciting fields in machine learning, provides an opportunity to automatically dig out the forecasting patterns hidden in the data. This is a new way to obtain the effective forecasting patterns, rather than extracting the man-made physical parameters from the active region.

The current work is focused on how to automatically mine forecasting patterns from solar observational data with a deep learning algorithm. The continuous and consistent solar observations of the photospheric magnetic field provided by the Michelson Doppler Imager on board the *Solar and Heliospheric Observatory* (*SOHO*/MDI) and the Helioseismic and Magnetic Imager on board the *Solar Dynamics Observatory* (*SDO*/HMI) make it possible to set up a deep learning based solar flare forecasting model. We collect line-of-sight magnetograms of active regions with few selection criteria, and fuse the *SDO*/HMI data with *SOHO*/MDI data to generate a big data set from 1996 April to 2015 October. Based on this data set, a solar flare forecasting model supported by a deep learning algorithm is developed, and its performances are comprehensively tested.

This article is organized as follows: The data is described in Section 2. A forecasting model based on the deep learning method is proposed in Section 3. The performance of the proposed model is evaluated in Section 4, and, finally, discussions and conclusions are provided in Section 5.

## 2. Data

In order to collect the continuous and consistent observational data to support the forecasting model, we employ the observational data from *SOHO*/MDI (Scherrer et al. 1995) and *SDO*/HMI (Schou et al. 2012).

### 2.1. Selection of Active Regions

A solar flare occurs when the stored magnetic energy is suddenly released in the solar corona, while the magnetic field at the site of the energy release cannot be accurately measured, and the observed photospheric magnetic fields in active regions are taken as the data set for modeling. *SOHO*/MDI and *SDO*/HMI provide a continuous and high-quality photospheric magnetic field observation. *SOHO*/MDI began its routine observation on 1996 January 1 and terminated on 2011 April 12. *SDO*/HMI, which is the successor to *SOHO*/MDI, began its routine observation on 2010 April 30. Line-of-sight magnetograms of active regions can be obtained from the tracked active region patch data product (Bobra et al. 2014) on the Joint Science Operations Center (JSOC) database (http://jsoc.stanford.edu/ajax/lookdata.html). For MDI data, line-of-sight magnetograms of active regions can be obtained from the MDI Tracked Active Region Patches (*mdi.Mtarp*). The MDI active region data from 1996 April 23 to 2010 April 30 are included. The cadence of MDI Tracked Active Region Patches is 96 minutes. For HMI data, line-of-sight magnetograms of active regions can be obtained from the HMI Active Region Patch (*hmi.Mharp_720s*). The HMI active region data from 2010 May 01 to 2015 October 01 are included. The cadence of the HMI Active Region Patch is 12 minutes. The magnetogram of an active region shows the magnetic flux distribution on the photosphere in this area.

The temporal cadence and spatial resolution of HMI data are higher than that of MDI data. In order to fuse the HMI data with the MDI data, the temporal cadence of HMI data is down-sampled to coincide with that of MDI data. For the spatial resolution, Liu et al. (2012) proposed a method to convert the HMI data to the MDI proxy data by comparing the HMI and MDI data on a pixel-by-pixel basis. We use the same method to convert the HMI data to the MDI proxy data. First, we reduce the spatial resolution of HMI data to be comparable to that of the MDI data by convolving a two-dimensional Gaussian function. The full width at half maximum of the Gaussian function is 4.7 HMI pixels, and the Gaussian function is truncated by 15 HMI pixels. Second, in order to obtain the same pixel size, the HMI pixels are averaged to the MDI pixels. Finally, we correct the pixel value using the equation $MDI = -0.18 + 1.40 * HMI$ (Liu et al. 2012) to generate the MDI proxy data.

The tracked active region patch data product, which can provide magnetograms in patches for active regions in their entire lifetime, captures more patches than the NOAA active region. Hence, the tracked active region patches with the faint magnetic field often have no designated NOAA active region numbers. These tracked patches without NOAA correspondence are excluded in the present work. Furthermore, Cui et al. (2007) estimates the influence of projection effects on the solar flare productivity, and finds that projection effects can be ignored for the active regions located within the $\pm 30°$ of the solar disk center. Therefore, only magnetograms of active regions with central longitudes located within $\pm 30°$ of the central meridian are included to avoid the influence of projection effects.

### 2.2. Definition of Flaring and Nonflaring Samples

The importance of solar flares is classified as A, B, C, M, or X according to the peak magnitude of the soft X-ray flux

**Table 1**
Sample Sizes with Different Forecasting Periods and Flare Importance Thresholds for MDI Data

| Forecasting Periods | Flare Thresholds | $N_f$ | $N_q$ | $\frac{N_f}{N_q}$ |
|---|---|---|---|---|
| 6 hr | C | 9165 | 108114 | 8.5% |
| | M | 1975 | 115304 | 1.7% |
| | X | 181 | 117098 | 0.2% |
| 12 hr | C | 11405 | 102874 | 14.0% |
| | M | 3857 | 113422 | 3.4% |
| | X | 383 | 116896 | 0.3% |
| 24 hr | C | 21151 | 96128 | 22.0% |
| | M | 6904 | 110375 | 6.3% |
| | X | 890 | 116389 | 0.8% |
| 48 hr | C | 29190 | 88089 | 33.1% |
| | M | 11526 | 105753 | 10.9% |
| | X | 1845 | 115434 | 1.6% |

**Note.** $N_f$ and $N_q$ stands for the number of flaring samples and the number of nonflaring samples, respectively.

**Table 2**
Sample Sizes with Different Forecasting Periods and Flare Importance Thresholds for HMI Data

| Forecasting Periods | Flare Thresholds | $N_f$ | $N_q$ | $\frac{N_f}{N_q}$ |
|---|---|---|---|---|
| 6 hr | C | 3051 | 45435 | 6.7% |
| | M | 364 | 48122 | 0.8% |
| | X | 33 | 48453 | 0.1% |
| 12 hr | C | 4882 | 43604 | 11.2% |
| | M | 674 | 47812 | 1.4% |
| | X | 67 | 48419 | 0.1% |
| 24 hr | C | 7355 | 41131 | 17.9% |
| | M | 1175 | 47311 | 2.5% |
| | X | 136 | 48350 | 0.3% |
| 48 hr | C | 10782 | 37704 | 28.6% |
| | M | 1992 | 46494 | 4.3% |
| | X | 237 | 48249 | 0.5% |

**Note.** $N_f$ and $N_q$ stand for the number of flaring samples and the number of nonflaring samples, respectively.

observed by the *Geostationary Operational Environmental Satellite* system. We use this classification of solar flares in the forecasting model. The solar flare data are downloaded from https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-features/solar-flares/x-rays/goes/xrs/. A magnetogram of an active region is considered to be a flaring sample when at least one given or greater flare happens within a given forecasting period starting at the observation of the magnetogram. Conversely, it is considered a nonflaring sample. The sample sizes with different forecasting periods and flare importance thresholds for the MDI and the HMI data are shown in Tables 1 and 2, respectively.

## 3. Method

Many physical parameters have been proposed to characterize the complexity and nonpotentiality of active regions, and they play a role such as hand-craft parameters in the standard feed-forward neural network (Figure 1(a)). The hand-craft parameters of the standard feed-forward neural network only contain part of the information artificially extracted from magnetograms. Instead of manually extracting the physical parameters, the convolutional neural network, which is widely used to process image data, can automatically extract flaring patterns from a large number of magnetograms (Figure 1(b)).

There are four main types of layers in a convolutional neural network:

1. Convolutional layer. The convolutional layer performs the convolution between the input image and the filter defined by the convolutional layer. The convolutional layer is used to extract forecasting patterns from input images.
2. Nonlinear layer. The nonlinear layer provides the nonlinear transform after the convolutional layer. The rectified linear unit ($f(x) = \max(0, x)$) is applied in the layer. The nonlinearity is added into the neural network by this layer.
3. Pooling layer. The pooling layer works on statistics concerning the mean or maximum value within a sliding region. The pooling layer reduces the dimensionality of parameters in the network and keeps their important information.

4. Fully connected layer. Neurons in the fully connected layer are connected to all neurons in the previous layer. This layer is the last layer of the convolutional neural network. High-level reasoning is done in this layer.

Comprising these basic layers, we set up the convolutional neural network for the solar flare forecasting (Figure 2). Parameters of the convolutional neural network can be learned from the data set by using the iterative gradient based optimization. However, some hyperparameters should be set before the optimization process (Bengio 2012). Usually, hyperparameters are set by experience. There are two types of hyperparameters: model hyperparameters, which specify the structure of the convolutional neural network, and training hyperparameters, which determine how the convolutional neural network is trained.

The main model hyperparameters include the following:

1. Size of the neural network. A large neural network can fit the training data well, but more training data is required to optimize parameters of the model. With the limitation of the data size, the number of layers should be controlled. In our recent network, we have five layers, which consist of two convolutional layers and three fully connected layers. For each layer, the size of the layer should be specified. We set 64 filters in each convolutional layer to ensure that there are sufficient filters to extract forecasting patterns from the data. And, according to the experience of filter design, the kernel size of the filter is set to $11 \times 11$.
2. Weight initialization. Weights are initialized by using a zero-mean Gaussian with a 0.02 standard deviation. The gradient descent based optimization cannot assure to find the global optimization solution, hence different random weights could get different results. However, the optimization method can find the suboptimal solutions that can obtain similar results for the different initializations of the random weights.
3. Preprocessing input data. The standard convolutional neural network requires a fixed size for all input images. Input images of arbitrary sizes are transformed to the fixed size via cropping or warping (He et al. 2014). But the cropped region may not contain the entire active region, hence we warp the input image to $100 \times 100$ pixels.
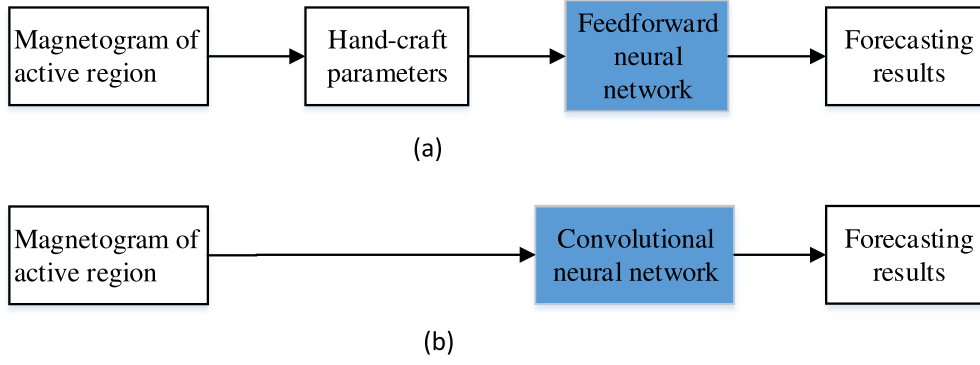
**Figure 1.** Difference between standard feed-forward neural network and convolutional neural network in solar flare forecasting. Components in shaded boxes need to be learned from the data set. (a) Standard feed-forward neural network for solar flare forecasting. (b) Convolution neural network for solar flare forecasting.
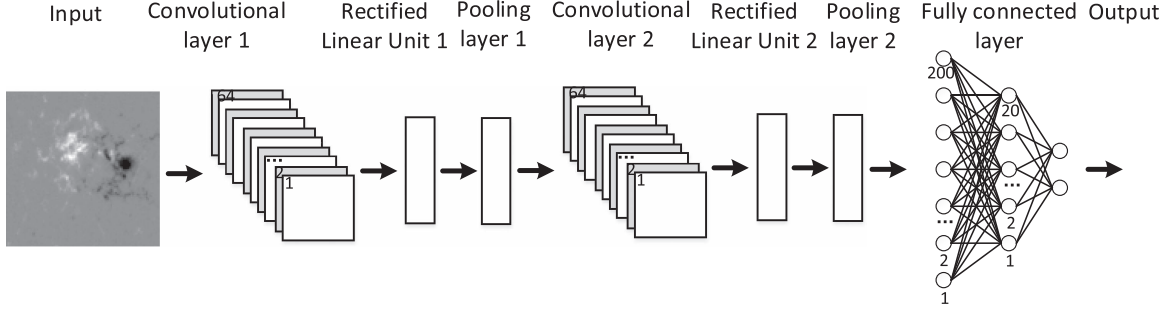


**Figure 2.** Structure of convolutional neural network for solar flare forecasting.

When the neural network works, magnetograms of active regions are warped. And a warped magnetogram is directly fed into the first convolutional layer. This layer performs the convolution on the input data with 64 $11 \times 11$ filters. After the convolutional layer, the rectified linear unit is applied to add the nonlinear transform into the network. Then the max pooling is operated to overcome the overfitting problem of the network. The first convolutional layer, nonlinear layer, and pooling layer are repeated to construct the second convolutional layer, nonlinear layer, and pooling layer, respectively. These layers are used to extract forecasting patterns from magnetograms of active regions. These patterns act as inputs for the three fully connected layers, and then fully connected layers are used to divide magnetograms of active regions into two classes, one stands for the flaring forecast and the other for the nonflaring forecast.

Once the architecture of the convolutional neural network for solar flare forecasting is determined, weights of different layers in the network should be learned from the data. These network parameters can be learned by using the stochastic gradient descent method. In this method, the network's parameters $\theta$ are updated by using gradient descent on a subset of the training data.

$$\theta^{(t)} = \theta^{(t-1)} - \alpha \frac{1}{N} \sum_{n=1}^{N} \frac{\partial L(Z_n, \theta)}{\partial \theta}, \quad (1)$$

where $t$ is the training step, $\alpha$ is the learning rate, $N$ is the number of samples in this training step, $Z_n$ is the nth sample in the training set, and $L$ is the loss function.

The main training hyperparameters include the following:

1. Learning rate. The learning rate ($\alpha$) determines how quickly the gradient updates follow the gradient direction.

If the learning rate is very small, the model will converge too slowly; if the learning rate is too large, the model will diverge. Here $\alpha$ is set to 0.01.

2. Loss function. The loss function ($L$) compares the network's output for a training example against the intended ground truth output.

$$L = -\sum_i y_i \log(z_i), \quad (2)$$

where $y_i$ is the output of the $i$th network output unit, and $z_i$ is the $i$th value of the target output.

After we set these training hyperparameters, the convolutional neural network can be trained as follows:

1. Forward propagation. A magnetogram of an active region is taken as an input going through the forward propagation, and then the output of the neural network is obtained. The loss function is estimated.
2. Back propagation. Gradients of the loss function with respect to all weights are back propagated through the network. Values of filter weights in the convolutional layer and connection weights in the fully connected layer are updated by the gradient descent to minimize the loss function.
3. Iteration. The forward propagation and back propagation are repeated until the stop criterion is satisfied. In order to prevent overfitting, the early stopping strategy, which stops training once performance on the testing set stops increasing, is often used in practice.

When the training process is finished, the performance of the forecasting model can be estimated by the testing data. In the testing process, testing magnetograms of active regions are input into the trained convolutional neural network, the network

|  | Forecasting Flaring | Forecasting Nonflaring |
| --- | --- | --- |
| Observed Flaring | True Positive (TP) | False Negative (FN) |
| Observed Nonflaring | False Positive (FP) | True Negative (TN) |

would go through the forward propagation and output the forecasting results. Comparing forecasting results to actual results, the performance of the forecasting model can be estimated.

# 4. Results

## 4.1. Performance Indexes

There are two types of outputs in the solar flare forecasting model. One is flaring forecast and the other is the nonflaring forecast. In order to quantificationally estimate the performance of this binary forecasting model, the contingency table is defined in Table 3.

As shown in Table 3, there are four possible outcomes for a binary forecasting model. The flaring sample is considered to be a positive class, and the nonflaring sample is considered to be a negative class. The number of samples that are correctly forecasted as "positive" are true positive (TP), while the number of samples that are correctly forecasted as "negative" are true negative (TN). The number of samples that are wrongly forecasted as "positive" are false positive (FP), while the number of samples that are wrongly forecasted as "negative" are false negative (FN).

As shown in Tables 1 and 2, the number of flaring samples is far less than the number of nonflaring samples, so this is a class imbalance problem (Bloomfield et al. 2012; Huang et al. 2012; Bobra & Couvidat 2015). In this problem, the performance can be estimated for positive classes or negative classes independently. We can also use a skill score that normalizes the performance of a model to a reference forecast, to take into account the effect of the class distribution.

Four metrics, true positive rate (TP rate), true negative rate (TN rate), false positive rate (FN rate), and false negative rate (FN rate), are defined to measure the performance of positive class and negative classes, respectively. The TP rate is the percentage of positive instances correctly classified.

$$\text{TP rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3}$$

TN rate is the percentage of negative instances correctly classified.

$$\text{TN rate} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{4}$$

FN rate is the percentage of negative instances that are misclassified.

$$\text{FP rate} = \frac{\text{FP}}{\text{TN} + \text{FP}} = 1 - \text{TN rate}. \tag{5}$$

FN rate is the percentage of positive instances that are misclassified.

$$\text{FN rate} = \frac{\text{FN}}{\text{TP} + \text{FN}} = 1 - \text{TP rate}. \tag{6}$$

Two skill scores, Heidke skill score (HSS) and true skill score (TSS), are used to estimate the performance of models for both classes. HSS measures the accuracy of the correct forecast relative to the random forecast.

$$\text{HSS} = \frac{\frac{\text{TP} + \text{TN}}{N} - \frac{(\text{TP} + \text{FP})(\text{TP} + \text{FN}) + (\text{TN} + \text{FN})(\text{TN} + \text{FP})}{N^2}}{1 - \frac{(\text{TP} + \text{FP})(\text{TP} + \text{FN}) + (\text{TN} + \text{FN})(\text{TN} + \text{FP})}{N^2}}, \tag{7}$$

where $N = \text{TP} + \text{TN} + \text{FP} + \text{FN}$.

$$\text{TSS} = \text{TP rate} - \text{FP rate}. \tag{8}$$

TSS is the difference between the TP rate and FP rate.

## 4.2. Performance Evaluations

The data set should be partitioned into the training set and the testing set. The training set is used to learn parameters of the solar flare forecasting model, and the performance of the forecasting model is evaluated on the testing data. We can evaluate the performance of a forecasting model in two ways:

1. Considering the temporal component inherent in the forecasting model and the possible difference between the MDI data and the MDI proxy data converted from the HMI data, we split all the MDI data into the training set, and the remaining MDI proxy data is treated as the testing data.

2. As usual, tenfold cross-validation is used to estimate the forecasting result and its uncertainty. The data set is randomly partitioned into 10 mutually exclusive subsets, and then one fold is used as the testing data and the remaining nine folds are used as the training data. This process is repeated 10 times until all of these folds are tested. Finally, the 10 results are averaged to produce a single estimation for the performance of a forecasting model.

### 4.2.1. Training on MDI Data and Testing on HMI Data

The solar flare forecasting model is trained and tested on the Caffe platform (Jia et al. 2014), which is a deep learning framework developed by the Berkeley Vision and Learning Center. In the training phase, weights in the forecasting model are iteratively adjusted to minimize the value of the training error on the MDI data. All samples in the training set are used once in one turn of training iterations. Taking the first convolutional layer in the neural network (Figure 2) as an example, we try to understand the performance of this network. The convolutional layer consists of 64 convolutional kernels of size 11 × 11. These convolutional kernels are initialized by the random values shown in Figure 3(a). During the training process, weights of the convolutional kernels become more and more orderly, as shown in Figures 3(b) and (c), and finally these weights stabilize after the iterations of 3000 steps shown in Figure 3(d). The animation shows the changing process of weights from initial random variables to stable forecasting patterns.
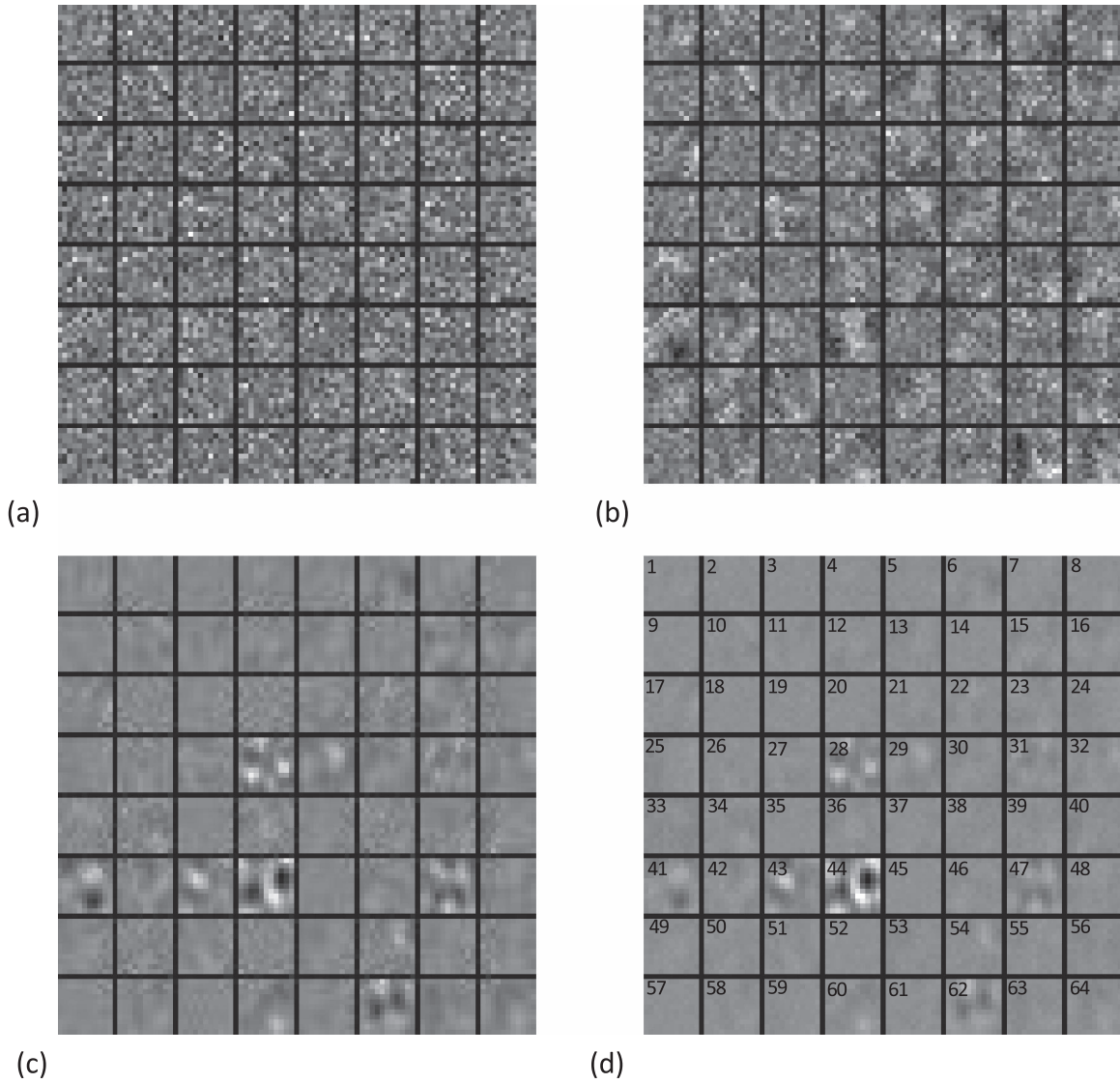
(a)



(b)



(c)



(d)

**Figure 3.** Variations of convolutional kernels in the first convolutional layer during different training iterations. (a) Initial random filter weights. (b) Filter weights after 1000 training iterations. (c) Filter weights after 2000 training iterations. (d) Filter weights after 3000 training iterations.

(An animation of this figure is available.)

In the testing phase, the performance of the forecasting model is evaluated on the MDI proxy data. Testing results of the training process are shown in Figure 4 for different flare importance thresholds within different forecasting periods. Figure 4 is a plot of the TP rate against the FN rate for different training iterations on the same testing set. Each point on the plot represents a TP rate and FN rate pair corresponding to a particular training iteration on the testing set. In each training iteration, all of the data in the training set is used to train the forecasting model. The point in the lower left corner or in the upper right corner means the forecasting model always outputs a nonflaring or flaring forecast, which is meaningless for the users. The top left corner in the plot represents the perfect forecasting model. The closer the point gets to the top left corner, the better the forecasting model performs. The performance of the forecasting model during overall training iterations is shown in this plot. Comparing results shown in Figure 4, we can find that in all of the given forecasting periods, the best performance of our model is for the flare with

importance over the X class, and the worst for the flare with importance over the C class.

As shown in Figure 3, weights of convolutional kernels in the convolutional neural network tend to be stable after 3000 training iterations, so the convolutional neural network with 3000 training iterations is considered to be the final forecasting model. The performances of the solar flare forecasting model for different flare importance thresholds within different forecasting periods are shown in Table 4. Ratios of the number of flaring samples to the number of nonflaring samples rely on flare importance thresholds and forecasting periods. As Bloomfield et al. (2012) mentioned, TSS is unbiased for the different sample ratio, we take TSS as a reasonable performance score for our model. This result is shown in Figure 5. From the distribution of TSS values in this figure, we can find that (1) the performance score increases when the flare importance threshold becomes larger; (2) the performance score does not change evidently in four given forecasting periods for the same flare importance threshold.

**Table 4**
Optimal Performances for Deep Learning Based Solar Flare Forecasting Model

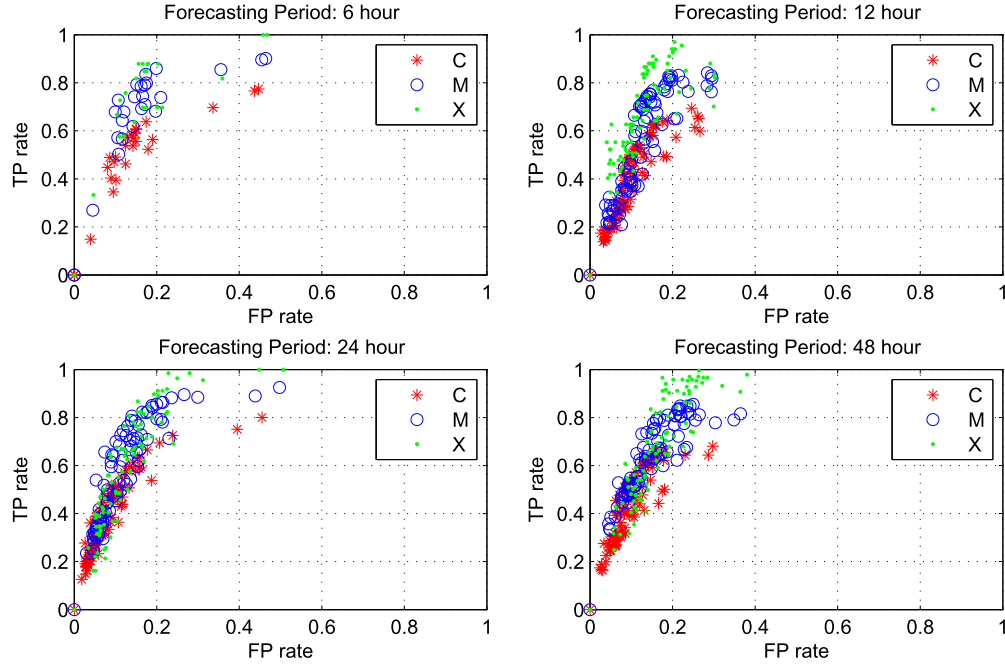| Forecasting Periods | Importance Thresholds | Contingency Table | | | | Evaluation Index | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | TP | FN | TN | FP | TP Rate | TN Rate | HSS | TSS |
| 6 hr | C | 1945 | 1106 | 37516 | 7919 | 0.64 | 0.83 | 0.227 | 0.463 |
| | M | 304 | 60 | 39782 | 8340 | 0.84 | 0.83 | 0.054 | 0.662 |
| | X | 29 | 4 | 40894 | 7559 | 0.88 | 0.84 | 0.006 | 0.723 |
| 12 hr | C | 3386 | 1496 | 32856 | 10748 | 0.69 | 0.75 | 0.243 | 0.447 |
| | M | 559 | 115 | 38388 | 9424 | 0.83 | 0.80 | 0.081 | 0.632 |
| | X | 61 | 6 | 41050 | 7369 | 0.91 | 0.85 | 0.014 | 0.758 |
| 24 hr | C | 5338 | 2017 | 31301 | 9830 | 0.73 | 0.76 | 0.339 | 0.487 |
| | M | 999 | 176 | 38398 | 8913 | 0.85 | 0.81 | 0.143 | 0.662 |
| | X | 118 | 18 | 40899 | 7451 | 0.87 | 0.85 | 0.025 | 0.714 |
| 48 hr | C | 7191 | 3591 | 31607 | 6097 | 0.67 | 0.84 | 0.467 | 0.505 |
| | M | 1614 | 378 | 37711 | 8783 | 0.81 | 0.81 | 0.206 | 0.621 |
| | X | 229 | 8 | 39791 | 8458 | 0.97 | 0.82 | 0.042 | 0.791 |



**Figure 4.** Testing results of the training process for different flare importance thresholds within different forecasting periods for the deep learning based solar flare forecasting model.
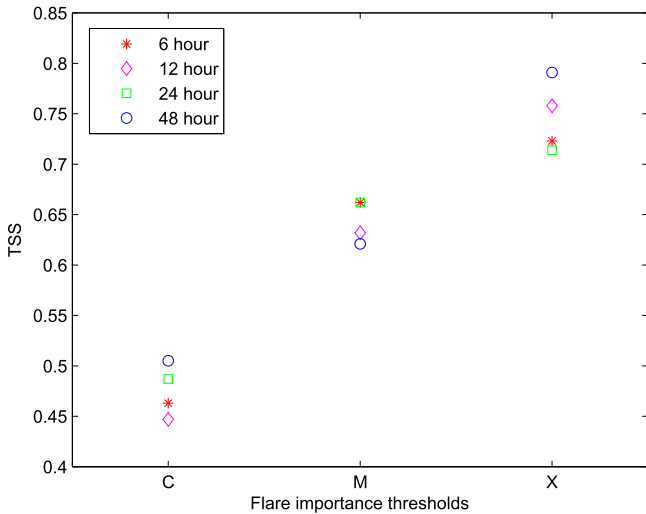


**Figure 5.** TSS of the deep learning based solar flare forecasting model for given flare importance thresholds within four forecasting periods.

In order to test the influence of noise on the forecasting results, we set 100 Gauss as the threshold of magnetic field in active regions. The magnetic field whose value is less than 100 Gauss in active regions is set to zero. The performances of the forecasting model are estimated by using the testing data with the 100 Gauss threshold and the raw testing data, respectively. Their performances are almost the same in the 24 hr flare forecasting (shown in Figure 6). This means that the deep learning based solar flare forecasting model is not sensitive for the weak magnetic field in the observed data. The noise levels of MDI and HMI are 26 G and 10 G respectively (Liu et al. 2012; Pietarila et al. 2013), which are much smaller than 100 G. Obviously, our model is not sensitive to the noise level.

This is our first time fusing the HMI data and MDI data together and forming a solar flare forecasting data set from 1996 April to 2015 October. The forecasting model is trained by using the MDI data and is tested by using the MDI proxy data, which means that the deep learning based flare forecasting model is suitable for different observational data from different satellites.
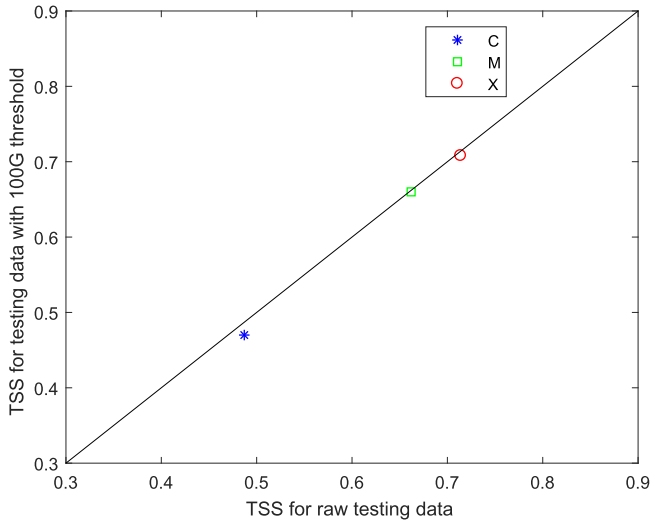
**Figure 6.** TSS for 24 hr flare forecasting models estimated by using the raw testing data and the testing data with a 100 G threshold.



**Figure 7.** Performance comparison of the deep learning based solar flare forecasting model between tenfold cross-validation (markers with error bars) and HMI data test (markers without error bars). In tenfold cross-validation, the markers show the mean of the 10 testing rounds and the error bars stand for their standard deviations.

### 4.2.2. Tenfold Cross-validation

In order to more statistically meaningfully estimate the performance of the forecasting model, the forecasting model is evaluated by the tenfold cross-validation method. The data set, which includes all the MDI and HMI samples, is randomly partitioned into 10 subsets. Therein, one subset is used as testing data and the remaining nine subsets are used as training data, this process is repeated 10 times, until each of the 10 subsets is used once as the testing data. The testing results are estimated by the mean and the standard deviation of these 10 rounds.

The performances of the deep learning based solar flare forecasting model estimated by tenfold cross-validation are shown in Figure 7. The performances tested by the HMI data are also plotted in the same figure to facilitate the comparison with the tenfold cross-validation testing results. The average TSS in tenfold cross-validation are little higher than the results tested by the HMI data. There are two possible reasons why better results are obtained for the tenfold cross-validation, one is because the training set for tenfold cross-validation is relatively larger than the MDI training data, and the other is because MDI data and HMI data are fused in the folded training set to get the more adaptable patterns for both MDI data and HMI data. In tenfold cross-validation, the uncertainty of the forecasting model can be estimated by the standard deviation of the 10 rounds. The testing results show that (1) as the flare threshold increases, the uncertainty of the forecasting model increases; (2) for the forecasting model with the same flare threshold, the uncertainty of the 6 and 12 hr flare forecasting model is larger than that of the 24 and 48 hr flare forecasting model. A possible reason for this is that there are more flaring samples for the smaller flare threshold with the longer forecasting period. For a deep learning method, more training samples mean that the forecasting model can be adjusted to be more optimized, and thus it performs better.

### 4.2.3. Comparison with Other Forecasting Models

Most of the forecasting models are trained and tested on different data sets with different definitions of events, forecasting frequencies, and forecasting periods. Taking 24 hr M flare forecasting as an example, the forecasting model
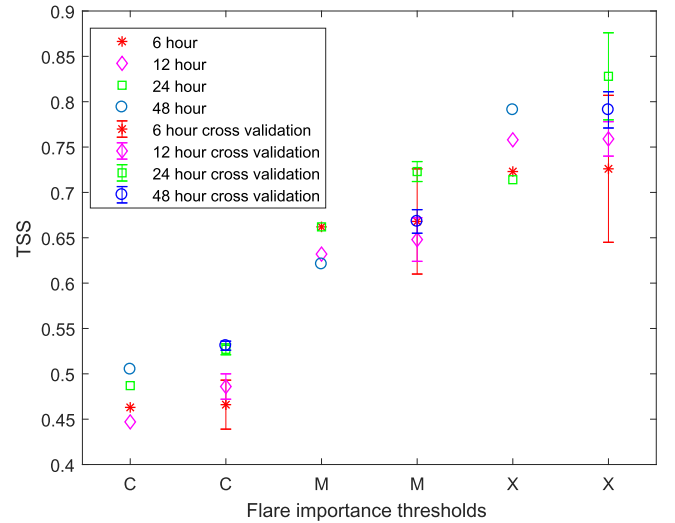
introduced by Murray et al. (2017) is carried out every 6 hr to output the 24 hr full-disk forecasts for the M class solar flares. The testing data is from 2015 to 2016, including 141 positive samples and 1489 negative samples. The forecasting model proposed by Muranushi et al. (2015) provides the 24 hr full-disk forecasts for the M flares with 1 hr cadence. The data set includes 1574 positive samples and 6837 negative samples from 2011 to 2012. The cross-validation is used to test the performance of the forecasting model. Bobra & Couvidat (2015) build a 24 hr active region based M flare forecasting model. The data set includes 303 positive samples and 5000 randomly selected negative samples, therein 30% of the data is used as testing data. The forecasting models introduced in Murray et al. (2017), Muranushi et al. (2015), and this work provide a rolling forecast. However, Bobra & Couvidat (2015) limits the time for flaring samples, which is exactly 24 hr prior to the peak time of solar flares. In this situation, Muranushi et al. (2015) and this work get the highest TP rate, Bobra & Couvidat (2015) has the highest TN rate and TSS.

Up to now, there has been no common data set to compare the performance of the different solar flare forecasting models. Because of the lack of common testing data, it is difficult to quantitatively compare the performance of the different forecasting models (Barnes et al. 2016). We list the forecasting results of four different models in Table 5 and find that the performance of our forecasting model is comparable with the others. However, it should be emphasized that the duration of total magnetograms continuously spans 19.5 years. In the model testing, the larger number of samples and the wider span of the testing data mean that the testing result is more sufficient and reliable.

## 5. Discussions and Conclusions

A deep learning based solar flare forecasting model is built to automatically extract forecasting patterns from magnetograms of active regions and forecast solar flares with the threshold of C, M, or X levels within the forecasting period of 6, 12, 24, or 48 hr. In order to produce an operational forecasting model

**Table 5**
Performance Comparisons of Four 24 hr M Flare Forecasting Models

| Reference | Definitions of Events | Forecasting Frequency | Forecasting Periods | Testing Samples | Evaluation Index | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| | | | | | TP Rate | TN Rate | TSS |
| Murray et al. (2017) | Full disk M class | Every 6 hr | 24 hr | 2015–2016 | 80% | 72% | 0.525 |
| Muranushi et al. (2015) | Full disk M class | Every 1 hr | 24 hr | 2011–2012 Cross-validation | 85% | 67% | 0.517 |
| Bobra & Couvidat (2015) | Active region M class | 24h prior to flare | 24 hr | 2010–2014 30% for testing | 83% | 93% | 0.765 |
| This work | Active region M class | Every 1.5 hr | 24 hr | 2010–2015 | 85% | 81% | 0.662 |
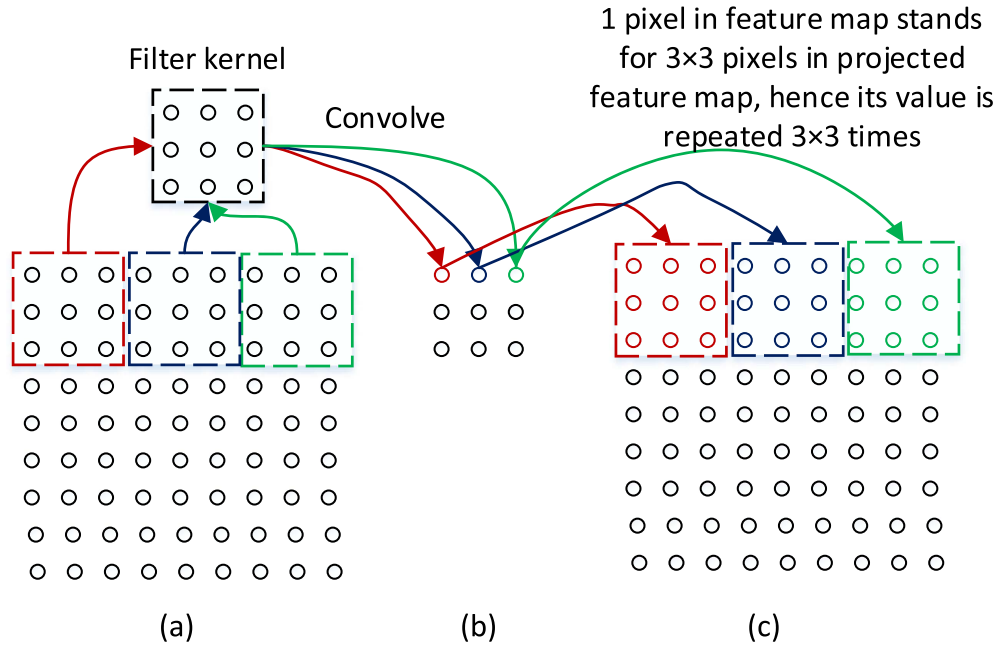


**Figure 8.** Schematic drawing shows how to generate the feature map and the projected feature map from the input image. (a) Input image. (b) Feature map. (c) Projected feature map.

from the observational data and evaluate its performances, we generate a data set including all active regions within $\pm 30°$ of the central meridian for line-of-sight magnetograms observed by the MDI and the HMI from 1996 April to 2015 October. The testing results of the forecasting model show that (1) some stable forecasting patterns can be learned from the MDI data by using the convolutional neural network, and these forecasting patterns can be applied to the HMI data; furthermore, the learned forecasting patterns are not sensitive to the noise in the observed data; (2) the deep learning based solar flare forecasting model works stationarily for the given forecasting periods (6, 12, 24, or 48 hr); (3) the performance of the proposed forecasting model is comparable to the other three state-of-the-art flare forecasting models.

The effectiveness of the deep learning based solar flare forecasting model is verified. The following question focuses on why this model can work well. In a convolutional neural network, the convolutional layer is used to extract patterns from input images. Figure 3(d) shows forecasting patterns extracted from magnetograms of active regions by the 64

$11 \times 11$ filters in the first convolutional layer of the convolutional neural network. We found that the weights in the most filters tend to 0, they are useless for the flare forecasting. By contrast, the stable nonzero filter weights are considered to be useful forecasting patterns, for example, the 44th and the 62nd filters in Figure 3(d). In order to analyze the physical implications of these forecasting patterns, we select line-of-sight magnetograms from two flare-productive active regions (AR 11158 and AR 10720) as the input data. The convolutional neural network requires a fixed size input image. However, this restriction comes from the fully connected layer, not the convolutional layer (He et al. 2014). Hence, convolutional filters can be applied to images with any sizes. The magnetogram of the AR 11158 is taken by the HMI and is transformed to the MDI proxy data. The magnetogram of the AR 10720 is taken by the MDI. The input data is convoluted with a filter to generate a feature map. A $k \times k$ filter with s stride is applied to $w \times h$ input data, we get a $\left(\frac{w-k}{s}+1\right) \times \left(\frac{h-k}{s}+1\right)$ feature map, and the feature map can be mapped into the same size with the input image to get
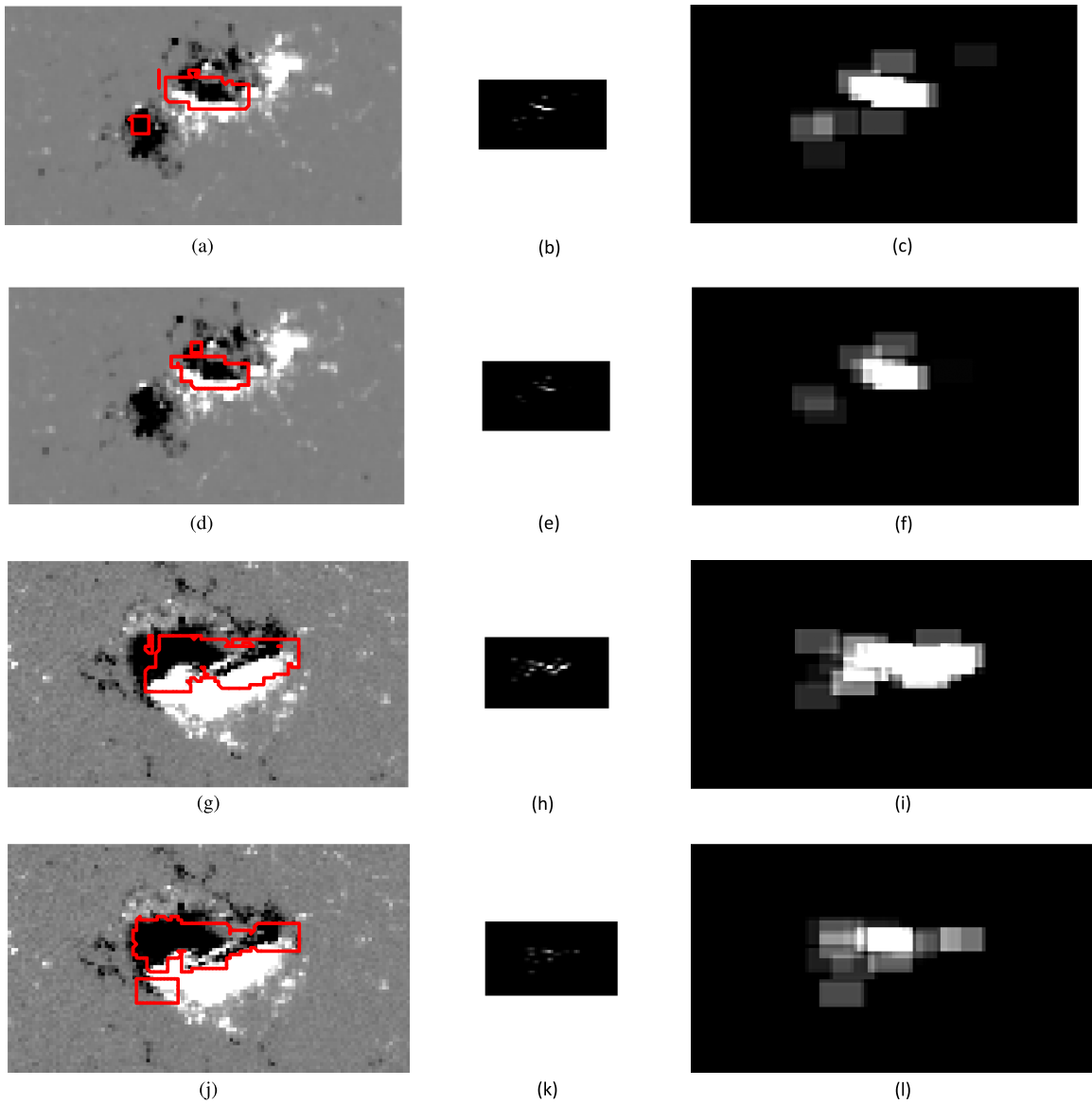
**Figure 9.** Magnetograms, feature maps, and projected feature maps for active regions 11158 and 10720. Line-of-sight magnetograms of AR 11158 and AR 10720 observed at 16:00 UT on 2011 February 13 and at 00:00 UT on 2005 January 15, respectively. The image is saturated at $\pm500$ Gauss. (a) Magnetogram of AR 11158 overlaid with contours of its 44th projected feature map. (b) 44th feature map of AR 11158. (c) 44th projected feature map of AR 11158. (d) Magnetogram of AR 11158 overlaid with contours of its 62nd projected feature map. (e) 62nd feature map of AR 11158. (f) 62nd projected feature map of AR 11158. (g) Magnetogram of AR 10720 overlaid with contours of its 44th projected feature map. (h) 44th feature map of AR 10720. (i) 44th projected feature map of AR 10720. (j) Magnetogram of AR 10720 overlaid with contours of its 62nd projected feature map. (k) 62nd feature map of AR 10720. (l) 62nd projected feature map of AR 10720.

the projected feature map. Taking the $3 \times 3$ filter with 3 stride as an example (Figure 8), the $9 \times 9$ input image is convolved with the $3 \times 3$ kernel with 3 stride to get a $3 \times 3$ feature map. In the feature map, 1 pixel stands for a $3 \times 3$ patch in the projected feature map, hence one value in the feature map is repeated $3 \times 3$ times to get the corresponding patch in the projected feature map.

Figure 9 shows the observed magnetograms, feature maps, and projected feature maps for AR 11158 and AR 10720. We apply convolution filters ($k = 11 \times 11$) with stride ($s = 3$) on the magnetogram of these two active regions ($110 \times 197$ for AR 10720 and $101 \times 185$ for AR 11158). We get a feature map ($34 \times 62$ for AR 10720 and $30 \times 58$ for AR 11158). One point in the feature map corresponds to an area with $11 \times 11$ points in the input magnetogram. Values at the most points of feature maps are close to zero, and only the highlighted areas

containing information related to solar flares, where the magnetic flux distribution is considered to be important for the solar flare forecasting. In order to locate highlighted areas of feature maps in their corresponding magnetograms, we overlay contours of projected feature maps on these magnetograms. It is found that our model pays attention to the area with the magnetic polarity-inversion line or the strong magnetic field in magnetograms of active regions.

Many studies have shed light on the relationship between the photospheric magnetic field and the solar flare. Cui et al. (2006) believe solar flares are closely related to the maximum horizontal gradient, the length of the neutral line in the photospheric magnetic field of active regions. Schrijver (2007) points out that major solar flares are associated with the magnetic flux emergence along the strong-field and high-gradient polarity-separation lines. Georgoulis & Rust (2007)

define the effective connected magnetic field strength that combines fluxes and field-line lengths between positive and negative fluxes; this metric is extremely correlated with the total unsigned magnetic flux (Barnes & Leka 2008). In short, solar flares preferentially occur in the complex active regions with strong field gradients and long polarity-inversion lines (Schrijver 2009). Therefore, the strong magnetic field, the high magnetic gradient, and the magnetic polarity-inversion line of active regions should be paid more attention for the solar flare forecasting. The region of interest in active regions determined by the convolutional neural network is consistent with the physical potential flaring area. This result implicates reasonability of the deep learning based solar flare forecasting model. Essential to this model is that the magnetic flux distribution in the region of interest is nonlinearly transformed and finally obtains some statistical relationships with solar flares.

The standard convolutional neural network requires a fixed size for all input images, hence we warp the input image to a fixed size. This means that the scaling information of images is ignored in our model. There are two ways to improve this problem: (1) record sizes of original images before they are warped and input these sizes into the forecasting model or (2) keep the original scale of images and expand them to the predefined size by supplementing the small random values or zero value. For example, we fix a large input size, the observational data of the small active region is placed in the middle of the image, and the remaining region is filled with the small random values or zero value. The impact of data preprocessing will be discussed in future work.

As the routine monitoring data of the Sun keeps getting bigger, the deep learning method, which can automatically extract forecasting patterns from the raw data, should play a more important role in the solar flare forecasting. The present work only extracts forecasting patterns from line-of-sight magnetograms of active regions. In the future, the forecasting patterns should be extracted from the the horizontal magnetic field and the extreme ultraviolet images of active regions. All the forecasting patterns extracted from the different observations can be compared and finally fused together to improve the performance of the solar flare forecasting.

## ORCID iDs

Huaning Wang https://orcid.org/0000-0003-2272-1469
Xinghua Dai https://orcid.org/0000-0003-1145-0253

## References

Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, SoPh, 283, 157
Arel, I., Rose, D. C., & Karnowski, T. P. 2010, IEEE Computational Intelligence Magazine, 5, 13
Barnes, G., & Leka, K. D. 2008, ApJL, 688, L107
Barnes, G., Leka, K. D., Schrijver, C. J., et al. 2016, ApJ, 829, 89
Barnes, J. P., Leka, K. D., Schumer, E. A., et al. 2007, SpWea, 5, S09002
Bengio, Y. 2012, eprint (arXiv:1206.5533)
Bloomfield, D. S., Higgins, P. A., & McAteer, R. T. J. 2012, ApJL, 747, L41
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Bornmann, P. L., & Shaw, D. 1994, SoPh, 150, 127
Boucheron, L. E., Al-Ghraibah, A., & McAteer, R. T. J. 2015, ApJ, 812, 51
Colak, T., & Qahwaji, R. 2009, SpWea, 7, S06001
Cui, Y. M., Li, R., Wang, H. N., et al. 2007, SoPh, 242, 1
Cui, Y. M., Li, R., Zhang, L. Y., et al. 2006, SoPh, 237, 45
Falconer, D. A. 2001, JGR, 106, 25185
Gallagher, P. T., Moon, Y. J., & Wang, H. M. 2002, SoPh, 209, 171
Georgoulis, M. K., & Rust, D. M. 2007, ApJL, 661, L109
Giovanelli, R. G. 1939, ApJ, 89, 555
Guerra, J. A., Pulkkinen, A., & Uritsky, V. M. 2015, SpWea, 13, 626
He, K., Zhang, X., Ren, S., & Sun, J. 2014, eprint (arXiv:1406.4729)
Hinton, G. E., & Salakhutdinov, R. R. 2006, Sci, 313, 504
Huang, X., & Wang, H. N. 2013, RAA, 13, 351
Huang, X., Wang, H. N., & Dai, X. H. 2012, SCPMA, 55, 1956
Huang, X., Yu, D. R., Hu, Q. H., et al. 2010, SoPh, 263, 175
Huang, X., Zhang, L. Y., Wang, H. N., et al. 2013, A&A, 549, A127
Jakimiec, M. 1987, AcA, 37, 271
Jia, Y. Q., Shelhamer, E., Donahue, J., et al. 2014, eprint (arXiv:1408.5093)
Korsós, M. B., Baranyi, T., & Ludmány, A. 2014, ApJ, 789, 107
Korsós, M. B., Ludmány, A., Erdelyi, R., et al. 2015, ApJL, 802, L21
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Lee, K., Moon, Y. J., Lee, J. Y., et al. 2012, SoPh, 281, 639
Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1296
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Li, R., Wang, H. N., He, H., et al. 2007, ChJAA, 7, 441
Liu, Y., Hoeksema, J. T., Scherrer, P. H., et al. 2012, SoPh, 279, 295
Mason, K. D., & Hoeksema, J. T. 2010, ApJ, 723, 634
McIntosh, P. S. 1990, SoPh, 125, 251
Miller, R. W. 1988, JRASC, 82, 191
Muranushi, T., Shibayama, T., Muranushi, Y. H., et al. 2015, SpWea, 13, 778
Murray, S. A., Bingham, S., Sharpe, M., et al. 2017, SpWea, 15, 577
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Pietarila, A., Bertello, L., Harvey, J. W., et al. 2013, SoPh, 282, 91
Qahwaji, R., & Colak, T. 2007, SoPh, 241, 195
Raboonik, A., Safari, H, Alipour, N., et al. 2016, ApJ, 834, 11
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, SoPh, 162, 129
Schmidhuber, J. 2015, NN, 61, 85
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, SoPh, 275, 229
Schrijver, C. J. 2007, ApJL, 655, L117
Schrijver, C. J. 2009, AdSpR, 43, 739
Song, H., Tan, C. Y., Jing, J., et al. 2009, SoPh, 254, 101
Wang, H. N., Cui, Y. M., Li, R., et al. 2008, AdSpR, 42, 1464
Wheatland, M. S. 2004, ApJ, 609, 1134
Yu, D. R., Huang, X., Hu, Q. H., et al. 2009, ApJ, 709, 321
Yu, D. R., Huang, X., Wang, H. N., et al. 2009, SoPh, 255, 91
Yu, D. R., Huang, X., Wang, H. N., et al. 2010, ApJ, 710, 869