Predicting Solar Flares with Machine Learning: Investigating Solar Cycle Dependence

Xiantong Wang, Yang Chen, Gabor Toth, Ward B. Manchester, Tamas I. Gombosi, Alfred O. Hero, Zhenbang Jiao, and Yang Liu

Abstract

A deep learning network, long short-term memory (LSTM), is used to predict whether an active region (AR) will produce a flare of class $\Gamma$ in the next 24 hr. We consider $\Gamma$ to be $\geq M$ (strong flare), $\geq C$ (medium flare), and $\geq A$ (any flare) class. The essence of using LSTM, which is a recurrent neural network, is its ability to capture temporal information on the data samples. The input features are time sequences of 20 magnetic parameters from the space weather Helioseismic and Magnetic Imager AR patches. We analyze ARs from 2010 June to 2018 December and their associated flares identified in the Geostationary Operational Environmental Satellite X-ray flare catalogs. Our results produce skill scores consistent with recently published results using LSTMs and are better than the previous results using a single time input. The skill scores from the model show statistically significant variation when different years of data are chosen for training and testing. In particular, 2015–2018 have better true skill statistic and Heidke skill scores for predicting $\geq C$ medium flares than 2011–2014, when the difference in flare occurrence rates is properly taken into account.

1. Introduction

Solar flares are energetic eruptions of electromagnetic radiation from the Sun lasting from minutes to hours. The terrestrial impact of small flares is limited, but strong flares have a significant impact on the upper atmosphere. Increased ionization affects the total electron content, which in turn affects radio wave propagation and global positioning system accuracy. Ionospheric heating causes the atmosphere to expand, increasing the mass density of and drag on satellites and altering their orbits. Strong flares are often accompanied by coronal mass ejections (CMEs) that can cause a substantial impact on the Earth environment. Therefore, it is very worthwhile to improve the prediction of solar flares, especially larger ones. During solar cycle 24, nearly 800 $M$ or $X$ flares were observed. While posing a significant threat, the rareness of extreme events and the complexity of the flares makes solar flare time and intensity predictions a very challenging task.

Although the triggering mechanism of solar flares and the factors determining the solar flare strength are far from being well understood, it is shown by multiple studies that solar flares are caused by the sudden release of free energy brought by magnetic reconnection in the coronal field. What has come to be known as the standard model for flares and CMEs (also called the CSHKP model; Camichael 1964; Sturrock 1966; Hirayama 1974; Kopp & Pneuman 1976), involves the rise of a sheared core or flux rope that results in magnetic reconnection in the surrounding arcade structure. Several variations of this model have been developed that incorporate different initiation mechanisms (Masuda et al. 1994; Forbes & Acton 1996; Manchester 2003; Török et al. 2004). A number of review papers summarize these works and many others (Green et al. 2018).

Since the photospheric magnetic field drives the coronal field, it is possible that the evolution patterns of the photospheric magnetic field may serve as indicators of the triggering process of flares and CMEs. Those features include the size of the active regions (ARs), the integrated magnetic flux, the integrated current helicity, the magnetic field gradient measurements, the shear angle of the magnetic field structure, and so on. The Helioseismic and Magnetic Imager (HMI) on the Solar Dynamics Observatory (SDO) satellite, launched a decade ago, has been providing high-cadence, high-resolution photospheric vector magnetic field observations since 2010. The space weather HMI AR patches (SHARPs; Bobra et al. 2014) contain time series data localized to individual ARs with many precalculated quantities based on the AR magnetic field. We will use these SHARP quantities to train our machine-learning model.

Machine learning, a subfield of artificial intelligence, utilizes past data as a “learning context” for the computer program that allow it to make predictions of the future state of the system. The advent of the Solar Heliophysics Observatory/Michelson Doppler Imager (MDI) and SDO/HMI missions provided sufficient data for machine-learning algorithms to predict solar activities. At first, the line-of-sight (LOS) component of the photospheric magnetic field measured by the MDI instrument was used by several groups to forecast solar flares using machine-learning algorithms (Song et al. 2009; Yu et al. 2009; Yuan et al. 2010; Ahmed et al. 2013; Huang et al. 2018). The support vector machine (SVM) algorithm was used by Boucheron et al. (2015) for a classification task on time series of MDI data from 2000 to 2010. However, the LOS magnetic field component does not include all of the magnetic field information, so later studies used the vector magnetic field data once it became available from the HMI instrument. Bobra & Couvidat (2015) used the SVM trained with SHARP parameters for AR classification tasks. Nishizuka et al. (2018) built a residual deep neural network using not only parameterized photospheric magnetograms but also chromospheric images. Jonas et al. (2018) used observations from the photosphere, chromosphere, transition region, and corona as input for the machine-learning algorithm,
which gave a comparable result to the work done by Bobra & Couvidat (2015) and Nishizuka et al. (2018).

The machine-learning algorithms in these studies mentioned so far do not fully utilize the time dependence of the input. Among various kinds of machine-learning algorithms, recurrent neural networks (RNNs) are suitable to analyze time series input. Long short-term memory (LSTM; Hochreiter & Schmidhuber 1997; Gers et al. 2000) networks, a particular kind of RNN, have succeeded in many sequence classification and prediction tasks, including speech recognition, time series forecasting, handwriting recognition, and so on (Hastie et al. 2005; Graves et al. 2013). Most recently, Chen et al. (2019), Jiao et al. (2019) and Liu et al. (2019) used LSTMs on SHARP parameters, which achieved better performance for predicting solar flares compared to previous works. Sun et al. (2019) identified key signals for strong flares from SHARP parameters using time series clustering on LSTM predictions.

In this paper, we apply the LSTM algorithm on the SHARP parameters from the SDO/HMI vector magnetic field to predict the maximum solar flare class produced by an AR in the next 24 hr. The inputs are 24 hr time series of SHARP parameters with 12 minutes cadence. The observations of ARs are time sequences; hence, LSTMs are suitable for this kind of input. First, our results show consistency with recently published work by Liu et al. (2019) that also uses the LSTM algorithm. Second, we also find that the skill scores vary substantially when using different years of ARs in the training and testing set. This indicates that data samples should be carefully chosen for the model evaluation. It is also of interest to understand in what respect these years differ in the solar cycle from each other that may make the solar flare prediction less or more successful.

The rest of this paper is organized as follows. Section 2 describes how we collect data and build the training and testing sets. Section 3 describes the LSTM architecture we are using in this work. Section 4 explains the metrics used to evaluate the model performance. Section 5 shows the results of this study and compares them with previous work. The solar cycle dependence of the prediction skills is also presented in this section. Section 6 describes our conclusions.

### 2. Details of the Data Preparation

#### 2.1. Data Set

We use the SHARP summary parameters as the input data of the prediction model. SHARPs are a data product derived from vector magnetograms taken from the HMI on board the SDO (Bobra et al. 2014). The summary parameters are calculated based on the HMI AR patches (HARPs), which are rectangular boxes surrounding the ARs that are moving with the solar rotation and the evolution of the ARs. Table 1 lists the 20 key parameters used in this work. The SHARP summary parameters are downloaded for all ARs from the Joint Science Operations Center (http://jsoc.stanford.edu) from 2011 to 2018. The solar flare events are identified from the NOAA Geostationary Operational Environmental Satellite (GOES) flare list (Garcia 1994). In the GOES flare list, flare events are listed with class, start, end, and peak intensity times of each event. The peak time of the flare events are assigned as the “event time” when constructing the data samples. The numbers of ARs and flare events in different years are summarized in Table 2. Note that C flares outnumber the A and B flares, suggesting that most of the A and B flares are missed when their relatively weak signal falls below the X-ray background.

The input of our model are the 24 hr long time series (that we call sequences) extracted from the full time series of SHARP summary parameters of ARs. To guarantee the quality of the data, some time sequences are dropped, especially when the ARs are at the limb. The criteria for dropping unqualified time sequences are as follows.

1. In order to avoid projection effects, the longitude of the HARP region center is within the range of ±68° from the central meridian.
2. The fraction of missing frames in a time sequence has to be less than 5%.

| Parameter | Description |
|-----------|-------------|
| USFLUX | Total unsigned flux, in Mx |
| MEANGAM | Mean inclination angle, gamma, in deg |
| MEANGBT | Mean value of the total field gradient, in G Mm⁻¹ |
| MEANGBZ | Mean value of the vertical field gradient, in G Mm⁻¹ |
| MEANGBH | Mean value of the horizontal field gradient, in G Mm⁻¹ |
| MEANJZD | Mean vertical current density, in mA m⁻² |
| TOTUSIZ | Total unsigned vertical current, in A |
| MEANALP | Total twist parameter, alpha, in 1 Mm⁻² |
| MEANIZH | Mean current helicity, in G² m⁻¹ |
| TOTUSHI | Total unsigned current helicity, in G² m⁻¹ |
| ABSNIZH | Absolute value of net current helicity, in G² m⁻¹ |
| SAVNCP | Sum of the absolute value of the net currents per polarity, in A |
| MEANPOT | Mean photospheric excess magnetic energy density, in ergs cm⁻² |
| TOTPOT | Total photospheric magnetic energy density, in ergs cm⁻² |
| MEANSHR | Mean shear angle (measured using $B_{\text{mean}}$), in deg |
| SHRGT45 | Percentage of pixels with a mean shear angle greater than 45°, in % |
| SIZE | Projected area of patch on image in microhemisphere |
| SIZE,ACR | Projected area of active pixels on image in microhemisphere |
| NACR | Number of active pixels in patch |
| NPIX | Number of pixels within patch |
3. The starting times of two time sequences are separated by 1 hr.

The target value (or label) of each sequence is the maximum flare class produced by the AR in the next 24 hr after the end time of the sequence. The NOAA AR number is used to match the HARP and AR numbers in the GOES flare list. However, while GOES flares are strictly identified with NOAA ARs, we note that a single AR may be split among multiple HARPs or that a HARP may contain multiple ARs. Consequently, we find that 20% of HARPs have this mismatch issue, which may lead to a potential error when we assign the maximum value of the feature

\[ \text{fl}_i \]

to the next 24 hr after the end time. We therefore use the NOAA AR number to match the GOES ARs that may be missed or improperly attributed to the HARP.

Because the various SHARP features have different scales and units, the original data samples are normalized before input into the machine-learning model. Let \( z^n_i \) denote the normalized value of the \( i \)th feature in the \( n \)th data sample, then

\[ z^n_i = \frac{v^n_i - \mu_i}{\sigma_i}, \]

where \( v^n_i \) is the original value of the feature \( i \) in the data sample \( n \), while \( \mu_i \) and \( \sigma_i \) are the mean and standard deviation of the feature \( i \) calculated from the entire data set, respectively.

### 2.2. Training/Testing Splitting

In order to properly assess the performance of the machine-learning algorithms, we need to split the samples (time sequences of SHARP summary parameters and corresponding maximum flare classes) into a training set and a testing set. The training set is used for training the machine-learning model, while the testing set is for assessing the prediction capability of the model. In the training process, the model learns from the input data and adjusts its parameters to fit the ground truth. Both variable selection and parameter estimation are included in this process. The samples in the testing set should be totally separated from the training set, otherwise there will be an artificial gain of performance, since the information in the training set is leaked to the testing set (Kaufman et al. 2012; Schutt & O’Neil 2013). Hence, separating the samples based on ARs is necessary to guarantee that sequences from one AR will not occur in both training and testing sets simultaneously. All of the training/testing splitting in this paper is conducted based on HARPs.

### 3. Architecture of Machine-learning Model

The RNN is a category of neural networks that can make use of sequential information (Pearlmutter 1989). This architecture is naturally used in solar flare prediction because the ARs evolve with time and the occurrence of the solar flares is most likely related to the time-dependent evolution of ARs. The RNNs are called recurrent because they perform the same task for every input from the sequence, but the output depends on the previous computations. Among the various RNN structures, the LSTM network is one of the most commonly used types of RNN (Hochreiter & Schmidhuber 1997). The LSTM networks are explicitly designed to avoid the long-term dependency problem, which is a major shortcoming for simpler RNNs. The key to LSTMs is a new cell state variable in the network, which is passed through the whole chain with only minor linear interactions. This allows the information to affect the results at a much earlier time, which mimics “long-term” memory.

The structure of the LSTM unit and network used in this work is shown in Figure 1. The left panel shows a single LSTM unit. Each unit takes an input vector \( x(t) \) consisting of the input features at a certain time point \( t \); the hidden state \( h(t-1) \) and cell state \( c(t-1) \) are from the previous LSTM unit. The right panel shows the structure of the two-layer LSTM network. The LSTM units in the second layer are the same as in the first layer, but their input vectors \( x_2(t) \) are the output vectors of the first-layer LSTM units. The relationships between the unit input, output, and internal states are given by the following
The Astrophysical Journal, 895:3 (13pp), 2020 May 20

Wang et al.

introduce the nonlinearity into the neural network. The weight matrices and bias vectors are trainable parameters, which are used by Liu et al. (2019) in their cost function; and $h$ is the hidden dimension of the LSTM unit, which is a hyperparameter in the model that reflects the model complexity, and we use 16 in this work. Here $c^{(l)} \in \mathbb{R}^h$ is the cell state vector, and there is only a linear relationship between the output and input cell states in a single LSTM unit. The cell state and hidden state vectors are passed to the next LSTM unit in the same layer. The output vectors in the first layer are taken as input in the second layer. The output vector of the last LSTM unit in the second layer is multiplied by an $h$ vector and passed to a sigmoid function to generate the final prediction value. The tanh and the sigmoid function

$$
\sigma(x) = \frac{1}{1 + e^{-x}}
$$

introduce the nonlinearity into the neural network. The weight matrices are applied to the input vectors, and $h$ is the hidden dimension of the LSTM unit, which is a hyperparameter in the model that reflects the model complexity, and we use 16 in this work. Here $i^{(l)}$, $f^{(l)}$, and $o^{(l)} \in \mathbb{R}^h$ are the activation vectors of the input, forget, and output gates, respectively; $g^{(l)} \in \mathbb{R}^d$ is an activation vector from the tanh function; and $h$ is the hidden dimension of the LSTM unit, which is a hyperparameter in the model that reflects the model complexity, and we use 16 in this work. Here $c^{(l)} \in \mathbb{R}^h$ is the cell state vector, and there is only a linear relationship between the output and input cell states in a single LSTM unit. The cell state and hidden state vectors are passed to the next LSTM unit in the same layer. The output vectors in the first layer are taken as input in the second layer. The output vector of the last LSTM unit in the second layer is multiplied by an $h$ vector and passed to a sigmoid function to generate the final prediction value. The tanh and the sigmoid function

$$
\sigma(x) = \frac{1}{1 + e^{-x}}
$$

The four quantities TN, TP, FN, and FP refer to the number of true-negative, true-positive, false-negative, and false-positive predictions, respectively. These four numbers can be combined to calculate the precision, recall (also known as probability of detection (POD)), false-alarm rate (FAR; also known as probability of false detection or false alarm (PFD or PFA)), true skill statistic (TSS), Heidke skill score (HSS), and accuracy (ACC), defined as

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

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

$$
\text{FAR} = \text{PFA} = \frac{FP}{FP + TN},
$$

$$
\text{TSS} = \frac{TP + FN}{TP + FN} - \frac{FP + TN}{FP + TN} = \text{POD} - \text{FAR},
$$

$$
\text{HSS} = \frac{2(\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN})}{(\text{TP} + \text{FN})(\text{FN} + \text{TN}) + (\text{TP} + \text{FP})(\text{FP} + \text{TN})}.
$$

We will use these quantities to evaluate the model performance. Precision and POD evaluate the model’s ability to identify positive events (1 being perfect, zero being worst),

\begin{table}[h]
\centering
\caption{The Ratio of Negative and Positive Samples $p_c$ in the Loss Function for Different Training Sets}
\begin{tabular}{cccccc}
\hline
Year & 2011 & 2012 & 2013 & 2014 & 2015 & 2015–18 \\
\hline
$\geq C$ & 3.91 & 4.08 & 4.27 & 4.46 & 4.35 & 4.35 \\
$\geq M$ & 29.41 & 30.30 & 31.25 & 34.48 & 34.48 & 34.48 \\
\hline
\end{tabular}
\end{table}
while FAR tests the model for correctly identifying negative events (zero being perfect, 1 being worst). The ACC, TSS, and HSS evaluate the overall skill, with the maximum value 1 being the perfect score. As we will see, the various skill scores have very different dependences on the fraction of positive and negative events in the training and testing sets. For an imbalanced data set, the ACC becomes less meaningful because the model’s output will be dominated by the majority of the data set. Artificial inflation will be caused to the POD (or the FAR) if the model is assigning all testing samples to be positive (or negative). However, in both cases, the model will not have any useful prediction skills. The TSS approaches the POD when the forecasting is dominated by correct forecasts of nonoccurrence, which is the case for solar POD when the forecasting is dominated by correct forecasts of nonoccurrence, which is the case for solar events. Dividing the equation by $P \cdot C$ results in a normalized expectation value:

$$e = \frac{\text{FAR} \cdot N}{P} - \frac{\text{POD} \cdot S}{C},$$  

If the ratios $S/C$ and $N/P$ are equal, then Equation (7) is proportional (with a negative sign) to the TSS, so in this special case, the TSS is optimal to evaluate and compare the usefulness of models with respect to minimizing the cost function. However, if $N/P$ is much larger than $S/C$, then a low FAR score is more important than maximizing the POD. We do not have good values for estimating $S$ and $C$, and in general, they vary with event types. Hence, there is no single skill score that can properly evaluate the forecasting performance, and one needs to be careful when models are compared. There are only four independent values (TN, TP, FN, and FP), and only their three ratios truly matter. This means that any three independent values defined above can be used. In practice, we concentrate on the POD, FAR, and HSS values, as these provide complete and intuitive information about the model’s performance. The TSS, while useful, is not an independent skill score, as it is simply the difference of the POD and FAR.

### 5. Results

#### 5.1. Training Process

The LSTM network is implemented in Python with the PyTorch package. PyTorch was originally a tensor calculation package for GPU, and the autograd feature (Paszke et al. 2017) makes it suitable for machine-learning tasks. A minibatch strategy (Li et al. 2014; Bottou et al. 2018) is used for faster convergence during back-propagation. The Adam optimizer (Kingma & Ba 2014) is used with the learning rate set to 0.001, and the other parameters are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is 1000. The model is trained for multiple epochs on the training set. In each epoch, the model goes through the training samples once. The model is trained for six epochs to generate the results presented in Table 5 and Figures 3 and 4. We found no statistically significant improvement in the performance after six epochs. To account for the randomness due to the order of training samples and the initial values of the trainable parameters, we perform 20 independent runs with different random seeds to get robust results.

#### 5.2. Skill Scores for Solar Flare Prediction

The prediction results from the models in this work are summarized in Table 4. The contingency table shows the number of TP, TN, FP, and FN events. Notice that the mean values of 20 runs are reported here, so the results are not integers. The TN is large because of the imbalanced data set. The TP for predicting $\geq C$ flares is much larger than that for predicting $\geq M$ flares. This is the result of more $\geq C$ flares. The FP is also large, which indicates that the model overpredicts the occurrence of the flare.

The mean values of the skill scores for the 20 runs are reported in Table 5, together with results obtained by earlier work for comparison. From Table 5, the skill scores POD, TSS, and FAR are better for predicting $\geq M$ flares than $\geq C$ flares,
Figure 2. Distribution of the number of recorded flares for classes B, C, M, and X from 2011–2018 in the GOES data set. Note that the number of B flares is much smaller than the number of C flares.

5.3. Comparison with Previous Results

In this subsection, we will compare the results of this paper to previously published works. In the past decade, there have been several works that applied machine learning–based models to predict solar flares. These models include multilayer perceptron (MLP; Florios et al. 2018), SVM (Qahwaj & Colak 2007; Yuan et al. 2010; Bobra & Couvidat 2015; Boucheron et al. 2015; Muranushi et al. 2015; Florios et al. 2018), deep flare net (DeFN; Nishizuka et al. 2018), and a recently published work (Liu et al. 2019) that also used the LSTM network.

The skill scores for different models are shown in Table 5. Even though the input and the testing samples may be different in those works using the MLP, SVM, and DeFN models, the results are representative of the best performance those models can achieve on the solar flare prediction task. The results from the LSTM models outperform the MLP, SVM, and DeFN models substantially on the HSS and FAR. The TSS from the DeFN is higher than that from the LSTM model; this is because of the high POD produced by the DeFN. Meanwhile, the DeFN produces a much higher FAR, which reduces the reliability of the positive predictions from the DeFN. The difference of the skill scores indicates that taking time series data into account can improve solar flare forecasting. As discussed in Section 5.4, we should use similar testing sets when citing the various models, since different train/test splitting can produce different skill scores. The notations for the LSTM models are defined as follows: LSTM-15_18 and LSTM-2015 are the models used in this paper, where LSTM-15_18 uses ARs from 2015–2018 for testing and LSTM-2015 uses ARs from 2015 for testing; LSTM-Liu uses the model reported in Liu et al. (2019), which uses ARs from 2015 for testing.

Several things can be noted about these evaluation metrics. The ACC is the least useful, since the data set is highly biased; hence, many weak flares that the model predicts are probably classified as false positives, which lowers the skill scores.

The skill scores for different models are shown in Table 5. Even though the input and the testing samples may be different in those works using the MLP, SVM, and DeFN models, the results are representative of the best performance those models can achieve on the solar flare prediction task. The results from the LSTM models outperform the MLP, SVM, and DeFN models substantially on the HSS and FAR. The TSS from the DeFN is higher than that from the LSTM model; this is because of the high POD produced by the DeFN. Meanwhile, the DeFN produces a much higher FAR, which reduces the reliability of the positive predictions from the DeFN. The difference of the skill scores indicates that taking time series data into account can improve solar flare forecasting. As discussed in Section 5.4, we should use similar testing sets when citing the various models, since different train/test splitting can produce different skill scores. The notations for the LSTM models are defined as follows: LSTM-15_18 and LSTM-2015 are the models used in this paper, where LSTM-15_18 uses ARs from 2015–2018 for testing and LSTM-2015 uses ARs from 2015 for testing; LSTM-Liu uses the model reported in Liu et al. (2019), which uses ARs from 2015 for testing. Several things can be noted about these evaluation metrics. (1) The ACC is the least useful since the data set is highly biased; flare events are rare, so the majority of samples in the testing set are negative. Predicting those negative samples correctly leads to a high accuracy; however, not much predictability for

while the precision and HSS show the opposite trend. The higher HSS is partially explained by the larger fraction of positive samples for ≥C flares (Doswell et al. 1990). However, essentially all skill scores for predicting any flares (≥A) are worse than predicting ≥C flares. The reason for this reduced performance is that A and B flares are not properly observed. The number of flares in different energy classes roughly obeys a power-law distribution (see Lu & Hamilton 1991 and Figure 8); thus, a large number of class B flares are missing in the GOES flare records (see Figure 2). The X-ray emission of many B flares can fall below the background emission level once an AR heats up, which causes those B flares to be unrecorded. Therefore, we are training and testing the model with mislabeled data samples for predicting any class of flares; hence, many weak flares that the model predicts are probably classified as false positives, which lowers the skill scores.

Table 5

Comparison of Skill Scores for Different Models

| Metric   | Model       | ≥M Class | ≥C Class | Any Class |
|----------|-------------|----------|----------|-----------|
| POD      | MLP         | 0.812    | 0.637    | ...       |
|          | SVM         | 0.692    | 0.746    | ...       |
|          | DeFN        | 0.947    | 0.809    | ...       |
|          | LSTM-Liu    | 0.881    | 0.762    | ...       |
|          | LSTM-2015   | 0.685    | 0.643    | 0.625     |
|          | LSTM-15_18  | 0.730    | 0.621    | 0.530     |
| Precision| MLP         | 0.143    | 0.451    | ...       |
|          | SVM         | 0.106    | 0.497    | ...       |
|          | DeFN        | 0.180    | 0.529    | ...       |
|          | LSTM-Liu    | 0.222    | 0.544    | ...       |
|          | LSTM-2015   | 0.311    | 0.677    | 0.670     |
|          | LSTM-15_18  | 0.282    | 0.635    | 0.702     |
| ACC      | MLP         | 0.855    | 0.778    | ...       |
|          | SVM         | 0.824    | 0.803    | ...       |
|          | DeFN        | 0.858    | 0.822    | ...       |
|          | LSTM-Liu    | 0.909    | 0.829    | ...       |
|          | LSTM-2015   | 0.929    | 0.858    | 0.814     |
|          | LSTM-15_18  | 0.945    | 0.883    | 0.800     |
| HSS      | MLP         | 0.204    | 0.389    | ...       |
|          | SVM         | 0.141    | 0.472    | ...       |
|          | DeFN        | 0.263    | 0.528    | ...       |
|          | LSTM-Liu    | 0.347    | 0.539    | ...       |
|          | LSTM-2015   | 0.394    | 0.567    | 0.519     |
|          | LSTM-15_18  | 0.382    | 0.557    | 0.473     |
| TSS      | MLP         | 0.669    | 0.449    | ...       |
|          | SVM         | 0.520    | 0.562    | ...       |
|          | DeFN        | 0.802    | 0.634    | ...       |
|          | LSTM-Liu    | 0.790    | 0.607    | ...       |
|          | LSTM-2015   | 0.623    | 0.559    | 0.509     |
|          | LSTM-15_18  | 0.681    | 0.553    | 0.439     |
| FAR      | MLP         | 0.143    | 0.188    | ...       |
|          | SVM         | 0.172    | 0.184    | ...       |
|          | DeFN        | 0.145    | 0.175    | ...       |
|          | LSTM-Liu    | 0.091    | 0.155    | ...       |
|          | LSTM-2015   | 0.062    | 0.084    | 0.116     |
|          | LSTM-15_18  | 0.049    | 0.068    | 0.092     |
| P/N      | DeFN        | 0.034    | 0.244    | ...       |
|          | LSTM-Liu    | 0.029    | 0.243    | ...       |
|          | LSTM-2015   | 0.040    | 0.269    | 0.371     |
|          | LSTM-15_18  | 0.026    | 0.189    | 0.403     |

Note. Lines with asterisks are results from this work. Here LSTM-2015 uses the same time period as DeFN, while LSTM-15_18 uses the same time period as LSTM-Liu for the testing data set. The number ratios of positive and negative samples are also reported in this table.
strong flare events may actually be achieved. (2) Precision and POD reflect the model’s ability to make positive predictions. Precision is the fraction of correctly predicted samples among all predicted positive samples, and POD is the fraction of correctly predicted positives among the actual positive samples in the testing set. Therefore, precision provides more useful information about the predictability of rare events, while the POD by itself is not representative of the predictive skills. From Table 5, our model produces precision scores within the same range as Liu et al. (2019) and is better than the previous works listed. (3) The HSS and TSS are often considered as good metrics for evaluating model predictability in binary classification tasks. However, the TSS benefits from the large number of correctly predicted negative samples (TN), so it is much higher than the HSS in all tests. The HSS lessens the importance of TNs, but it is more sensitive to the fraction of positive samples in the testing set: the value of the HSS will be higher if there are more positive samples in the testing set (which can be artificially achieved by creating a testing set that contains a larger fraction of positive samples than the actual data). Our model gives a better HSS than previous models (MLP, SVM, and DeFN) that do not use time sequences and also has similar performance as the recently published results (Liu et al. 2019) using LSTM, which validates the correctness of this work. (4) The FARs for predicting $\geq C$ and $\geq M$ flares are all less than 0.1, which means that more than 90% of the negative predictions from our model are correct. This contributes greatly to the high HSS values. Notice that both the POD and FAR in our results are lower than those from LSTM-Liu, and our model produces fewer false alarms.

The details of our and the Liu et al. (2019) LSTM models are different, but they obtain similar skill scores according to Table 5. This suggests that both LSTM models extract most of the useful information from the SHARP parameters, and further improvement will require using more information from the observation.

5.4. Choosing Different Testing Years

As described in Section 2.2, it is important to totally separate the training and testing samples. However, whether choosing different years of flares for testing can have different skill scores is still unclear, since the previous works all used data before 2015 for training and after 2015 for testing. In this subsection, we conduct the training–testing process on different combinations of training and testing years. In Figure 3, we present the box plots of skill scores for 20 independent runs with the testing samples being one of the years from 2011–2015, and the other four years are used for training. (As shown by Table 2, there were very few large flares from 2016–2018, so those years are not suitable for testing and would not contribute much to training either.)

Figure 3 shows that training on 2011–2014 and testing on 2015 gives the best HSS and FAR scores for predicting both $\geq C$ and $\geq M$ flares. The trend is different for the TSS for predicting $\geq M$ flares, because the TSS is dominated by the POD values, but this does not mean truly good prediction of rare events, such as $\geq M$ flares. Good prediction of rare events requires very few false alarms, and this will produce a high HSS. Apparently, the model is quite “restrained” on making positive predictions for the 2015 data, which improves its FAR and HSS scores. It is clear from Figure 3 that evaluating the model performance on different years will introduce significant differences in the results.

To investigate why the model produces fewer false alarms in 2015 than other years, we set up two linear regression models (LRMs) as the baseline. These two baseline models use the same training and testing samples as the LSTM model. The time sequences of the SHARP parameters are reshaped to one-dimensional vectors as the input of the first LRM, denoted by LRM-A. The elements of the reshaped one-dimensional vector are in the following order: $[f_1^t, ..., f_T^t, ..., f_1^n, ..., f_T^n]$, where $f_t^n$ is the value of the $n$th feature at time $t$. For the second LRM, denoted by LRM-B, the mean values of the time sequences are taken as input. Twenty independent runs are conducted, and the mean values of the skill scores from the LSTM model and two baseline models are shown in Figure 4. The difference between the LSTM model and LRM-A is the nonlinearity introduced by the LSTM network. The LRM-B eliminates the time sequence information and only inputs the average level of activity into the model. From Figure 4, it can be seen that (1) for predicting $\geq C$ flares, the LSTM gives the best HSS, followed by LRM-A and LRM-B. For predicting $\geq M$ flares, the LSTM model has a similar HSS as the LRM-A model, and both are better than the LRM-B model. This illustrates the importance of the time sequence information. (2) The LRMs give a larger POD and FAR than the LSTM. Therefore, the LSTM model has less tendency to make positive predictions, which results in a better HSS. The optimal case is when the model produces a high POD while also keeping a low FAR, which is not the case for an LSTM. This is the reason why an LSTM cannot provide a high TSS compared to LRMs.

All of the time sequence data taken by LRM-A result in better HSS, TSS, and FAR scores than LRM-B, which uses the average parameter values of the input sequence. This shows that even for simple linear models, time sequence information helps the model to make better predictions. Notice that the LSTM model gives a better HSS and FAR than the LRMs, while the TSS and POD are similar or even smaller than LRM-A. This means that the nonlinearity introduced by LSTM networks is effective to improve the HSS and FAR but not TSS and POD. All three models give the lowest FAR when the model is tested on 2015. This indicates that, regardless of the different algorithms, the temporal averaged SHARP parameters are relevant to the result of the FAR, as well as the time sequence parameters. In conclusion, the LSTM model produces more reliable positive predictions than the simple LRMs, although it misses more positive events. The model gives different results when being tested/trained on different years of data, apparently due to differences in the average SHARP parameters.

5.5. Training/Testing on Different Solar Cycle Phases

According to Section 5.4, training and testing on different years leads to very different results. One possible reason for the difference could be the changes in the data processing procedure. However, from Hoeksema et al. (2018), although the data processing techniques were modified in 2015 January, those changes will not have major effects on the data products used in this work.

To investigate if there are any intrinsic differences of the ARs from each year in the solar cycle, we first do the training and testing separately on each year. Since the number of ARs and flares is very small in 2016–2018, those three years are
grouped together. Four skill scores are collected to evaluate the model performance: HSS, TSS, POD, and FAR. The process for selecting data samples is as follows.

1. For the ARs in each year, randomly select 25% for testing and the rest of them for training.
2. Extract 24 hr time series of SHARP parameters for training and testing from the ARs selected in step 1 and label the time series with the maximum flare class in the next 24 hr from the GOES flare record.
3. Randomly drop negative samples in the training and testing sets to fix the ratio of positive and negative samples to be 0.05 for predicting \( \geq M \) flares and 0.3 for predicting \( \geq C \) flares.

Notice that in step 3, we fix the ratio of positive to negative samples to make the HSS directly comparable across different years. In addition, having a fixed positive/negative sample ratio makes the HSS and TSS behave similarly. We perform multiple runs with randomly dropping negative samples, so in fact, the runs use different data sets. For each run, the skill scores are the mean values of the model outputs from the third to the 10th epochs. This range of epochs is chosen based on the typical evolution of skill with epochs: there is an initial rapid improvement, followed...
by a plateau with random oscillations, and finally worsening trends due to overfitting. The averaging over multiple epochs reduces the random variation due to the relatively small data sets. In addition, the mean values better reflect the general performance of the model than picking the best epoch for each run.

The box plots of skill scores for predicting $\geq M$ flares are shown in Figure 5. Each box contains 100 data points from 100 runs using randomly selected ARs for training/testing and dropping negative samples in the data sets. Because there are few ARs and flare events during the three years from 2016–2018 (see Table 2), the skills scores are less centered and the number of outliers is larger than those from other years. The results show that training and testing on the data after 2015 produces better skill scores than the earlier years. The FAR has the most substantial difference for data sets after and before 2015, which is also the major reason for a better HSS and TSS,
since the POD does not vary much. We are not showing results for predicting $\geq M$ flares on each year separately because the $\geq M$ flares are too rare to give any statistically significant results on such small data samples.

The results in Figure 5 show that the prediction model has a better performance for 2015–2018 than for 2011–2014. To demonstrate this difference with better statistics, we conduct the training/testing process on two data sets using these two time periods and compare the results. The data-selecting strategy is the same as what we did for training/testing on each year, but now there are enough $M$ and $X$ flares to produce statistically robust results. The skill scores for predicting $\geq C$ and $\geq M$ flares are shown in Figures 6 and 7, respectively. For predicting $\geq C$ flares, the HSS and TSS scores are clearly and significantly better for 2015–2018 than for 2011–2014. The box plots of the POD overlap with each other, while the box plots of the FAR are well separated. This result indicates that for capturing $\geq C$ flares in the testing set, the model has a similar performance when trained on the data samples from two phases in a solar cycle. However, the model will produce many fewer false alarms if it is trained on data samples from the declining phase in a solar cycle. For predicting $\geq M$ flares, according to Figure 7, the difference of skill scores is less obvious. In general, the results from 2015–2018 show a larger variance because there are fewer $\geq M$ flares in this time range. The box plots of the HSS, TSS, and POD overlap to a large extent, except the model gives a higher FAR when trained/tested on 2015–2018.

There is a significant difference between 2011–2014 and 2015–2018 for predicting $\geq C$ flares but not $\geq M$ flares. A simple explanation could be different flare intensity distributions in two phases of a solar cycle. For example, if the frequencies of $C$ and $M$ flares are better separated from each other in one of the time periods (i.e., there are relatively fewer flares with energies near the $C/M$ class boundary), then it will help the model to correctly identify $\geq M$ flares. To check this possibility, we show the distribution of flare events as a function of energy on a log–log scale for the two time periods in Figure 8. The histograms are well approximated with straight lines, indicating that the flare intensity distributions approximately follow power laws (Lu & Hamilton 1991). While the amplitudes are different by about a factor of 3 (there are more flares during solar maximum than during solar minimum), the slopes of the two lines are very close to each other. So there is no obvious difference in the shape of the flare intensity distributions between the two phases of a solar cycle. There seems to be some excess of flares near the $C/M$ class boundary for 2015–2018, which may contribute to the worse performance on predicting $\geq M$ flares than $\geq C$ flares for this time period. However, we set the threshold of positive and negative classes to C8.0. All these excesses of flares are labeled as positive. The results show no difference in skill scores for the two time periods. Thus, we conclude that the difference of the skill scores in Figures 6 and 7 cannot be simply explained by different flare intensity distributions in the two time periods. A further possible explanation could be the coupled eruptions (or sympathetic flares) between different ARs (Moon et al. 2002; Titov et al. 2012). Sometimes one flare occurring in one AR leads to the occurrence of another flare(s) in other ARs. There are more ARs from 2011 to 2014, and it is anticipated that more sympathetic flares occur in this time range. The SHARP

![Figure 5. Box plots of skill scores for predicting $\geq C$ flares in different years from 2011–2018. The symbols in the figures are the same as in previous box plots.](image-url)
Figure 6. Box plots of skill scores for predicting $\geq C$ flares. The training/testing are conducted within one of two time periods: 2011–2014 and 2015–2018. The symbols in the figure are the same as in previous box plots.

Figure 7. Box plots of skill scores for predicting $\geq M$ flares. The training/testing are conducted within one of two time periods: 2011–2014 and 2015–2018. The symbols in the figure are the same as in previous box plots.
parameters may not contain enough information to predict sympathetic flares.

6. Conclusion

In this paper, we build a data set covering ARs from 2010–2018 from the Joint Science Operations Center (http://jsoc.stanford.edu). Each data sample is a time sequence of 20 SHARP parameters that represent the magnetic field properties of an AR. We develop an LSTM network to predict the maximum flare class $\Gamma$ in the next 24 hr produced by an AR. The prediction task is reduced to a binary classification when $\Gamma$ is a combination of classes above a certain threshold. We consider three different cases for $\Gamma$: $\geq M$, $\geq C$, and $\geq A$. The last case corresponds to predicting any flares. The training/testing splitting is based on ARs, which guarantees that the model is tested on data samples that it has never seen previously. The skill scores produced by the model vary substantially for different years, and we investigate the solar cycle dependence of the model performance. The main results of this paper are summarized as follows.

1. Flare prediction models should minimize the cost function. The cost function is not proportional and in general does not vary monotonically with the skill scores. This means that a comparison of model performance requires caution.

2. The LSTM-based model achieves a better HSS for predicting solar flares than previous approaches such as MLP, SVM, and DeFN. Using the time series information improves relevant skills. Our results are also comparable with the recently published work using a similar LSTM method.

3. Although more than 50% of the skill scores of the LSTM model can be acquired from simple LRMs, the non-linearity introduced by LSTM reduces the number of false alarms and improves the prediction skills of the model.

4. Previous works using AR data after 2015 for testing could introduce bias into the skill scores. If the model is trained on 2011–2014 and tested on 2015, it produces better skill scores than other combinations of training and testing years. This appears to be related to the difference of the time sequence--averaged values of the SHARP parameters used for the input, although the physical interpretation is not revealed. A possible way to avoid this issue is to mix ARs from different years in both the training and testing sets.

5. The skill scores from the model show a statistically significant variation when different years of data are chosen for training and testing. Here 2015–2018 has a better TSS, HSS, and FAR for predicting $\geq C$ flares than 2011–2014 when the difference in flare occurrence rates is properly taken into account.

Based on the results presented in this paper, the LSTM is a valid method for the solar flare prediction task. The skill scores from this paper are very close to those generated by other, different LSTM models, indicating that the information contained in the SHARP parameters is limited. In future work, we plan to use more observational information to further improve the flare prediction skills.

This work was primarily supported by NSF PRE-EVENTS grant No. 1663800 and NSF-DMS 1811083. We also acknowledge support from the NASA DRIVE Center at the University of Michigan under grant NASA 80NSSC20K0600. W.M. was supported by NASA grants 80NSSC18K1208, NNX16AL12G, and 80NSSC17K0686.

ORCID iDs

Xiantong Wang https://orcid.org/0000-0002-8963-7432
Gabor Toth https://orcid.org/0000-0001-8459-2100
Ward B. Manchester https://orcid.org/0000-0003-0472-9408
Tamas I. Gombosi https://orcid.org/0000-0001-9360-4951
Meng Jin https://orcid.org/0000-0002-9672-3873
Yang Liu https://orcid.org/0000-0002-0671-689X

References

Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, SoPh, 283, 157
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Bottou, L., Curtis, F. E., & Nocedal, J. 2018, SIAMR, 60, 223
Boucheron, L. E., Al-Ghraibah, A., & McAteer, R. T. J. 2015, ApJ, 812, 51
Carmichael, H. 1964, NASSP, 50, 451
Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, SpWea, 17, 1404
Doswell, C. A., III, Davies-Jones, R., & Keller, D. L. 1990, WiFor, 5, 576
Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, SoPh, 293, 28
Forbes, T., & Acton, L. 1996, ApJ, 459, 330
Garcia, H. A. 1994, SoPh, 154, 275
Gers, F. A., Schmidhuber, J., & Cummins, F. 2000, Neural Comput., 12, 2451
Graves, A., Mohamed, A.-r., & Hinton, G. 2013, in 2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (Piscataway, NJ: IEEE), 6645
Green, L. M., Török, T., Vršnak, B., Manchester, W., & Veronig, A. 2018, SSRv, 214, 46
Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. 2005, Math. Intell., 27, 83
Hirayama, T. 1974, SoPh, 34, 323
Hoeksema, J. T., Baldner, C. S., Bush, R. I., Schou, J., & Scherrer, P. H. 2018, SoPh, 293, 45
Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7
Jiao, Z., Sun, H., Wang, X., et al. 2019, arXiv:1912.06120
Jonas, E., Bobra, M., Shankar, V., Todd Hoeksema, J., & Recht, B. 2018, SoPh, 293, 48
Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. 2012, ACM Trans. Knowl. Discov. Data, 6, 15
Kingma, D. P., & Ba, J. 2014, arXiv:1412.6980
Kopp, R. A., & Pneuman, G. W. 1976, SoPh, 50, 85
