Draft version March 29, 2022
Typeset using LaTeX default style in AASTeX631

Predicting Solar Energetic Particles Using *SDO*/HMI Vector Magnetic Data Products and a Bidirectional LSTM Network

Yasser Abduallah,[1,2] Vania K. Jordanova,[3] Hao Liu,[1] Qin Li,[1,4,5] Jason T. L. Wang,[1,2] and Haimin Wang[1,4,5]

[1] *Institute for Space Weather Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA; ya54@njit.edu, wangj@njit.edu, haimin.wang@njit.edu*

[2] *Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA*

[3] *Space Science and Applications, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

[4] *Big Bear Solar Observatory, New Jersey Institute of Technology, 40386 North Shore Lane, Big Bear City, CA 92314, USA*

[5] *Center for Solar-Terrestrial Research, New Jersey Institute of Technology, University Heights, Newark, NJ 07102, USA*

ABSTRACT

Solar energetic particles (SEPs) are an essential source of space radiation, which are hazards for humans in space, spacecraft, and technology in general. In this paper we propose a deep learning method, specifically a bidirectional long short-term memory (biLSTM) network, to predict if an active region (AR) would produce an SEP event given that (i) the AR will produce an M- or X-class flare and a coronal mass ejection (CME) associated with the flare, or (ii) the AR will produce an M- or X-class flare regardless of whether or not the flare is associated with a CME. The data samples used in this study are collected from the *Geostationary Operational Environmental Satellite*'s X-ray flare catalogs provided by the National Centers for Environmental Information. We select M- and X-class flares with identified ARs in the catalogs for the period between 2010 and 2021, and find the associations of flares, CMEs and SEPs in the Space Weather Database of Notifications, Knowledge, Information during the same period. Each data sample contains physical parameters collected from the Helioseismic and Magnetic Imager on board the *Solar Dynamics Observatory*. Experimental results based on different performance metrics demonstrate that the proposed biLSTM network is better than related machine learning algorithms for the two SEP prediction tasks studied here. We also discuss extensions of our approach for probabilistic forecasting and calibration with empirical evaluation.

Keywords: Solar energetic particles; Coronal mass ejections; Solar flares; Solar activity

1. INTRODUCTION

Solar eruptions including flares and coronal mass ejections (CMEs) can endanger modern civilization. Solar flares are large bursts of radiation released into space; they appear as sudden and unexpected brightening in the solar atmosphere with a duration ranging from minutes to hours. CMEs are significant discharges of plasma and magnetic fields produced by the solar corona into the interplanetary medium (Lin & Forbes 2000). They are considered to be the largest-scale solar eruptions in the solar system and occur on a quasi-regular basis (Chen 2011; Webb & Howard 2012; Kilpua et al. 2017). Research shows that both flares and CMEs are magnetic events, sharing a similar physical process (Harrison 1995; Berkebile-Stoiser et al. 2012), though more work is performed to understand the correlation between them (Yashiro & Gopalswamy 2008; Kawabata et al. 2018). Large flares and accompanied CMEs cause solar energetic particles (SEPs). SEPs, composed of electrons, protons and heavy ions, are expedited by magnetic reconnection or shock waves associated with the CMEs (Brito et al. 2018; Huang et al. 2018). When SEP events are strong, they cause nuclear cascades in the Earth's upper atmosphere and also represent a radiation hazard to equipment in space that is not adequately protected (Reames et al. 2013; Jordanova et al. 2018; Roeder & Jordanova 2020).

Active regions (ARs), which manifest complex magnetic geometry and properties (Benz 2008), are the source of flares and CMEs (Chen 2011; van Driel-Gesztelyi & Green 2015). The lifetime of ARs ranges from days to months (van Driel-Gesztelyi & Green 2015). Recently, researchers combine machine learning with physical parameters derived from vector magnetograms provided by the Helioseismic and Magnetic Imager (HMI; Schou et al. 2012) on board the *Solar Dynamics Observatory* (*SDO*; Pesnell et al. 2012) to predict flares, CMEs, and SEPs. These physical parameters, including magnetic helicity and magnetic flux (Leka & Barnes 2003; Schrijver 2007; Moore et al. 2012), are part of the vector magnetic data products, named the Space-weather HMI Active Region Patches (SHARP; Bobra et al. 2014), produced by the *SDO*/HMI team.

Machine learning (ML) has been popular in predictive analytics for many years. ML is able to learn patterns from historical data and make predictions on unseen or future data (Alpaydin 2016; Goodfellow et al. 2016). For example, Liu et al. (2017) used random forests (RF) and the SHARP parameters to predict the occurrence of a certain class of flares in a given AR within 24 hours. Jonas et al. (2018) employed machine learning to extract relevant information from photospheric and coronal image data to perform flare prediction. Florios et al. (2018) adopted multiple machine learning algorithms including RF, multilayer perceptrons (MLP) and support vector machines (SVM) for flare forecasting. More recently, researchers started to use deep learning (DL), which is a branch of machine learning focusing on the use of deep neural networks, to enhance the learning outcome (Goodfellow et al. 2016). Huang et al. (2018) designed a convolutional neural network to learn patterns from line-of-sight magnetograms of ARs and used the patterns to forecast flares. Liu et al. (2019) adopted a long short-term memory (LSTM) network for flare prediction. Chen et al. (2019) employed LSTM and the SHARP parameters to identify solar flare precursors; the authors later extended their work by investigating solar cycle dependence (Wang et al. 2020). Similar ML and DL methods have been applied to CME and SEP forecasting. Bobra & Ilonidis (2016) used SVM to predict CMEs; Liu et al. (2020) extended their work by adopting recurrent neural networks including LSTM and gated recurrent units. Inceoglu et al. (2018) employed SVM and MLP to forecast if flares would be accompanied with CMEs and SEPs.

In this paper, we propose a new deep learning method, specifically a bidirectional long short-term memory (biLSTM) network, for SEP prediction using the *SDO*/HMI vector magnetic data products. We aim to solve two binary prediction problems: (i) predict whether an AR would produce an SEP event given that the AR will produce an M- or X-class flare and a CME associated with the flare (referred to as the FC_S problem); (ii) predict whether an AR would produce an SEP event given that the AR will produce an M- or X-class flare regardless of whether or not the flare is associated with a CME (referred to as the F_S problem). The proposed biLSTM is an extension of LSTM (Hochreiter & Schmidhuber 1997), both of which are well suited for time series forecasting (LeCun et al. 2015; Goodfellow et al. 2016). Unlike LSTM, which works in one direction, biLSTM works back and forth on the input data and then the patterns learned from the two directions are joined together to strengthen the learning outcome. In SEP prediction, the observations and physical parameters associated with ARs form time series, and hence biLSTM is suitable for our study.

The rest of this paper is organized as follows. Section 2 explains the data and data collection procedure used in our study. Section 3 describes our proposed deep learning method. Section 4 reports experimental results and discusses extensions of our approach for probabilistic forecasting and calibration. Section 5 concludes the paper.

2. DATA

In this work we adopt the Space-weather HMI Active Region Patches (SHARP; Bobra et al. 2014) that were produced by the *SDO*/HMI team and released at the end of 2012. These data are available for download, in the data series hmi.sharp, from the Joint Science Operations Center (JSOC).[1] The SHARP data provide physical parameters of active regions (ARs) that have been used to predict flares, CMEs and SEPs (Bobra & Ilonidis 2016; Liu et al. 2017; Inceoglu et al. 2018; Liu et al. 2019; Liu et al. 2020). We collected SHARP data samples from the data series, hmi.sharp_cea_720s, using the Python package SunPy (SunPy Community et al. 2015) at a cadence of 12 minutes. In collecting the data samples, we focused on the 18 physical parameters previously used for SEP prediction (Inceoglu et al. 2018). These 18 SHARP parameters include the absolute value of the net current helicity (ABSNJZH), area of strong field pixels in the active region (AREA_AC), mean characteristic twist parameter (MEANALP), mean angle of field from radial (MEANGAM), mean gradient of horizontal field (MEANGBH), mean gradient of total field (MEANGBT), mean gradient of vertical field (MEANGBZ), mean vertical current density (MEANJZD), mean current helicity (MEANJZH), mean photospheric magnetic free energy (MEANPOT), mean shear angle (MEANSHR), sum of flux near polarity inversion line (R_VALUE), sum of the modulus of the net current per polarity (SAVNCPP), fraction of area with shear $> 45°$ (SHRGT45), total photospheric magnetic free energy density (TOTPOT), total unsigned current helicity (TOTUSJH), total unsigned vertical current (TOTUSJZ), and total unsigned flux (USFLUX).

[1] http://jsoc.stanford.edu/

Since the 18 SHARP parameters have different units and scales, we normalized the parameter values using the min-max normalization procedure as done in Liu et al. (2020). Each data sample contains the 18 SHARP parameters. Let p_i^k be the original value of the ith parameter of the kth data sample. Let q_i^k be the normalized value of the ith parameter of the kth data sample. Let min_i be the minimum value of the ith parameter. Let max_i be the maximum value of the ith parameter. Then

$$q_i^k = \frac{p_i^k - min_i}{max_i - min_i}. \quad (1)$$

Appropriately labeling the data samples is crucial for machine learning. We surveyed M- and X-class flares that occurred between 2010 and 2021 with identified active regions in the *GOES* X-ray flare catalogs provided by the National Centers for Environmental Information (NCEI). As done in Bobra & Ilonidis (2016), we excluded ARs that were outside ± 70° of the central meridian because the SHARP parameters cannot be calculated correctly based on the vector magnetograms of the ARs that are near the limb due to foreshortening and projection effects.[2] We also excluded flares with an absolute value of the radial velocity of *SDO* being greater than 3500 m s^{-1}, low-quality HMI data as described by Hoeksema et al. (2014), and data samples with incomplete SHARP parameters. In this way we excluded data samples of low quality, and kept qualified data samples of high quality in our study. Furthermore, we collected and extracted information from NASA's Space Weather Database of Notifications, Knowledge, Information (DONKI)[3] to tag, for any given M- or X-class flare, whether it produced a CME and/or SEP event. We cross-checked the flare records in DONKI and *GOES* X-ray flare catalogs to ensure that each flare record was associated with an active region; otherwise the flare record was removed from our study.

We then created two databases of active regions (ARs) for the period between 2010 and 2021. ARs from 2010, 2016, and 2018-2021 were excluded from the study due to the lack of qualified data samples or the absence of SEP events associated with M-/X-class flares and CMEs. Thus, the databases contain ARs from six years, namely 2011-2015 and 2017. In our first database, referred to as the FC_S database, each record corresponds to an AR, contains an M- or X-class flare as well as a CME associated with the flare, and is tagged by whether the flare/CME produce an SEP event. In this database, there are 31 records tagged by "yes" indicating they are associated with SEP events while there are 97 records tagged by "no" indicating they are not associated with SEP events. In our second database, referred to as the F_S database, each record corresponds to an AR, contains an M- or X-class flare, and is tagged by whether the flare produces an SEP event regardless of whether or not the flare initiates a CME. In this database, there are 40 records tagged by "yes" indicating they are associated with SEP events while there are 700 records tagged by "no" indicating they are not associated with SEP events.

3. METHODOLOGY

3.1. *Prediction Tasks*

As mentioned in Section 1, we aim to solve the following two binary prediction problems. [**FC_S problem**] Given a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t and the flare initiates a CME, we predict whether x_t is positive or negative. Predicting x_t to be positive means that the AR will produce an SEP event associated with the flare/CME. Predicting x_t to be negative means that the AR will not produce an SEP event associated with the flare/CME. [**F_S problem**] Given a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t regardless of whether or not the flare initiates a CME, we predict whether x_t is positive or negative. Predicting x_t to be positive means that the AR will produce an SEP event associated with the flare. Predicting x_t to be negative means that the AR will not produce an SEP event associated with the flare. For both of the two binary prediction problems, we consider T ranging from 12 to 72 in 12-hour intervals as frequently considered in the literature (Ahmed et al. 2013; Bobra & Ilonidis 2016; Inceoglu et al. 2018; Liu et al. 2020).

In solving the two binary prediction problems, we first show how to collect and construct positive and negative data samples used in our study. Figure 1(a) (Figure 1(b), respectively) illustrates how to construct positive (negative, respectively) data samples for the FC_S problem where $T = 24$ hours. Refer to the FC_S database described in Section 2, which indicates whether a flaring AR that already produces an M- or X- class flare/CME will initiate an SEP event

[2] Notice that flaring ARs outside ± 70° of the central meridian may produce eruptive events that have increased probabilities to result in SEPs due to the magnetic connectivity with Earth. Excluding these flaring ARs may reduce the number of SEP events considered in the study. This is a limitation of our approach.

[3] http://kauai.ccmc.gsfc.nasa.gov/DONKI/

Table 1. Numbers of Positive and Negative Data Samples Constructed for Different Hours for the FC_S and F_S Problems Respectively

		12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
FC_S	Positive	994	2017	3055	4143	5221	6336
	Negative	2952	5522	7864	9976	11687	13135
F_S	Positive	1260	2561	3863	5207	6517	7864
	Negative	19593	31534	40619	48189	54718	59821

associated with the flare/CME. For the flaring AR, we collect data samples that are within the $T = 24$ hours prior to the peak time of the flare.

- If the flare/CME are associated with an SEP event, the data samples belong to the positive class and are colored (labeled) by blue as shown in Figure 1(a). Thus, for each blue (positive) data sample, there is an M- or X- class flare that is within the next 24 hours of the occurrence time of the data sample, the flare initiates a CME and the flare/CME are associated with an SEP event.
- If the flare/CME are not associated with an SEP event, the data samples belong to the negative class and are colored (labeled) by green as shown in Figure 1(b). Thus, for each green (negative) data sample, there is an M- or X- class flare that is within the next 24 hours of the occurrence time of the data sample, the flare initiates a CME but the flare/CME are not associated with an SEP event.

Constructing positive and negative data samples for the F_S problem is done similarly and its description is omitted.

Table 1 shows the numbers of positive and negative data samples constructed for the FC_S and F_S problems respectively. Consider the FC_S problem. The positive and negative data samples are constructed based on the 31 records tagged by "yes" and 97 records tagged by "no" in the FC_S database described in Section 2. When $T = 24$ hours and the cadence is 12 minutes, one would expect the total number of positive data samples to be 24 hr × 60 minutes/hr × (1/12 minutes) × 31 = 3720, and the total number of negative data samples to be 24 hr × 60 minutes/hr × (1/12 minutes) × 97 = 11640. However, the total number of positive (negative, respectively) data samples is 2017 (5522, respectively). This happens because we removed many data samples of low quality as described in Section 2. If a gap occurs in the middle of a time series due to the removal, we use a zero-padding strategy as done in Liu et al. (2020) to create a synthetic data sample to fill the gap. The synthetic data sample has zero values for all the 18 SHARP parameters. The synthetic data sample is added after the normalization of the SHARP parameter values, and hence the synthetic data sample does not affect the normalization procedure.

After explaining how to construct the positive and negative data samples, we now show how to solve the binary prediction problems. Consider again the FC_S problem where $T = 24$ hours. Here we want to predict whether a given test data sample x_t at time point t is positive (blue) or negative (green) given that there will be an M- or X-class flare within the next 24 hours of t and the flare initiates a CME. If there is an SEP event associated with the flare/CME, and we predict x_t to be positive (blue), then this is a correct prediction as illustrated in Figure 1(c). If there is an SEP event associated with the flare/CME, but we predict x_t to be negative (green), then this is a wrong prediction as illustrated in Figure 1(e). On the other hand, if there is no SEP event associated with the flare/CME, and we predict x_t to be negative (green), then this is a correct prediction as illustrated in Figure 1(d). If there is no SEP event associated with the flare/CME, but we predict x_t to be positive (blue), then this is a wrong prediction as illustrated in Figure 1(f). The F_S problem is solved similarly. In the following subsection we describe how to train our model and use the trained model to make predictions.

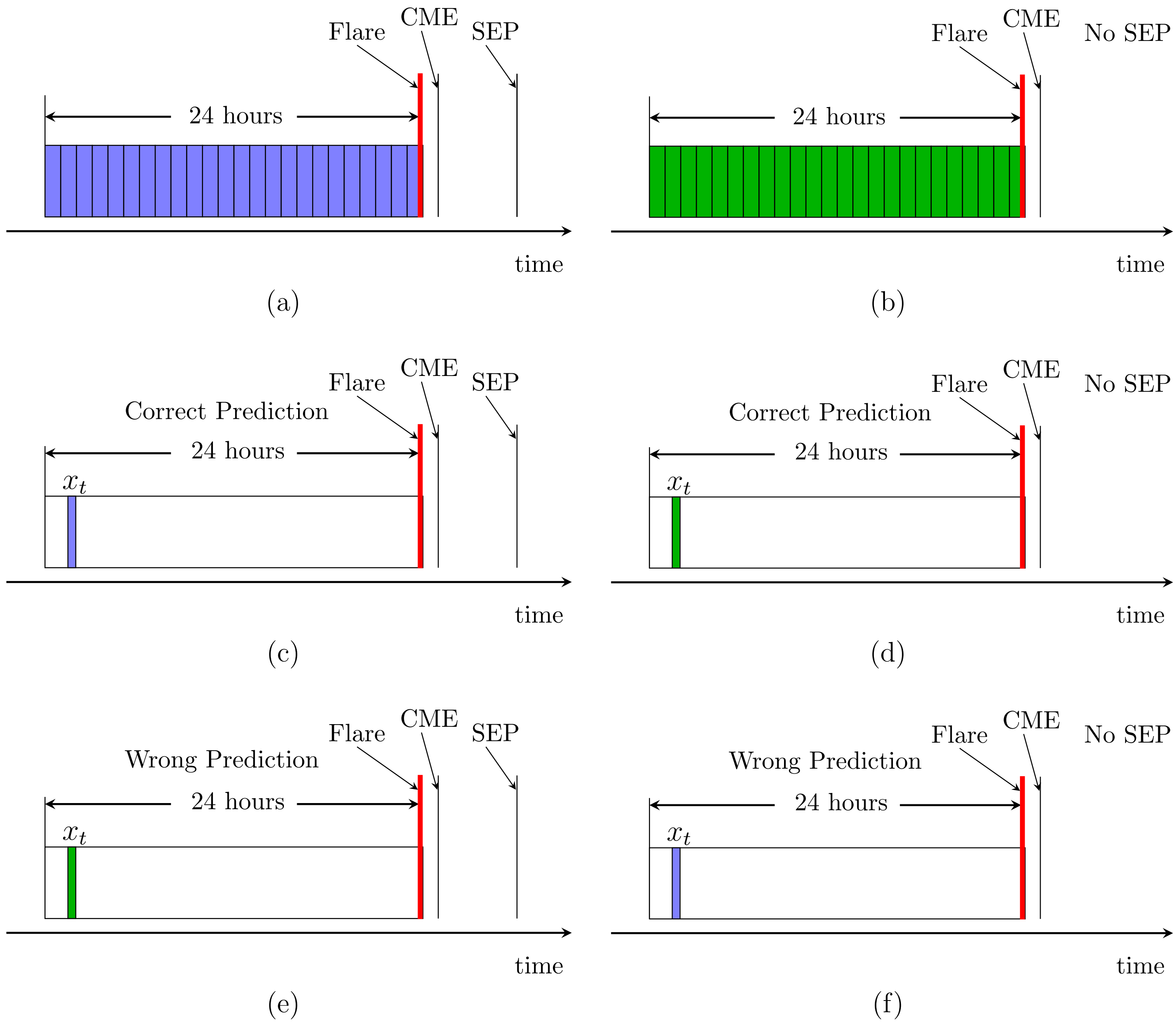


Figure 1. Collecting and constructing positive and negative data samples on a flaring AR for the FC_S problem where $T = 24$ hours and making predictions based on the collected data samples. The data samples are collected at a cadence of 12 minutes. Each rectangular box corresponds to 1 hour and contains 5 data samples. The red vertical line shows the peak time of an M- or X- class flare. (a) The blue rectangular boxes contain data samples that are within the 24 hours prior to the peak time of an M- or X- class flare that produces a CME and an SEP event; these blue data samples belong to the positive class. (b) The green rectangular boxes contain data samples that are within the 24 hours prior to the peak time of an M- or X- class flare that produces a CME but no SEP event; these green data samples belong to the negative class. (c) Illustration of a correct prediction for a test data sample x_t that is predicted to be positive. (d) Illustration of a correct prediction for a test data sample x_t that is predicted to be negative. (e) Illustration of a wrong prediction for a test data sample x_t that is predicted to be negative. (f) Illustration of a wrong prediction for a test data sample x_t that is predicted to be positive.

3.2. *Prediction Method*

We consider one of recurrent neural networks (RNNs) that is called long short-term memory (LSTM; Hochreiter & Schmidhuber 1997; Goodfellow et al. 2016) to build our model. LSTM has shown good results in solar eruption prediction (Liu et al. 2019; Liu et al. 2020). We create a model using bidirectional LSTM (biLSTM). Generally, a bidirectional RNN (Schuster & Paliwal 1997) functions by duplicating the initial recurrent layer in the network to obtain two layers so that one layer uses the input as is and the other duplicated layer uses the input in a reverse

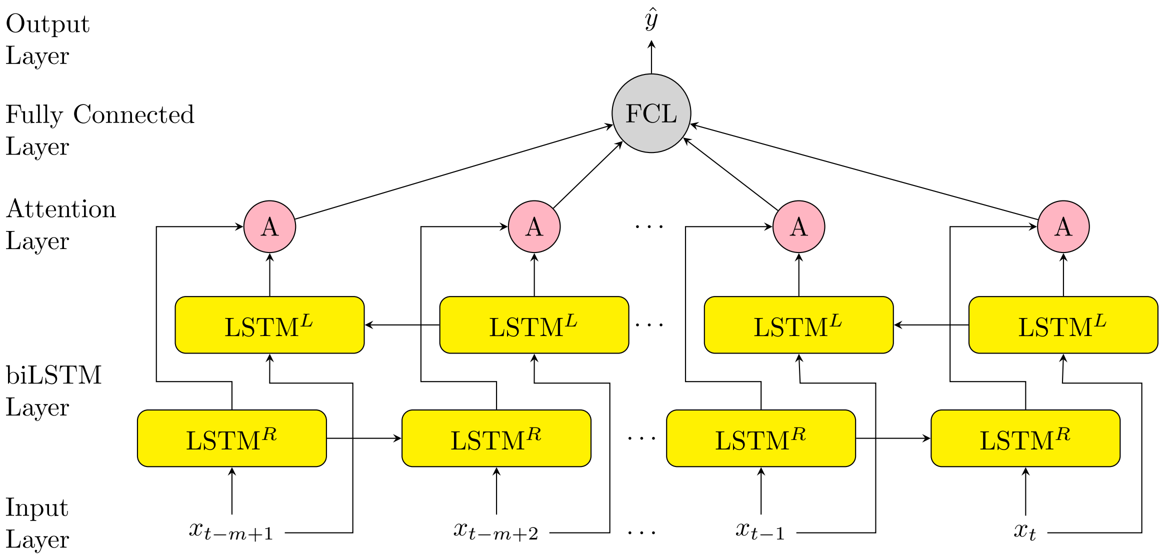


Figure 2. Architecture of the proposed biLSTM network. Yellow boxes represent biLSTM cells. These cells are connected to an attention layer (A) that contains m neurons, which are connected to a fully connected layer (FCL). (In the study presented here, m is set to 10.) During testing/prediction, the input to the network is a test data sequence with m consecutive data samples x_{t-m+1}, $x_{t-m+2} \ldots x_{t-1}$, x_t where x_t is the test data sample at time point t. The trained biLSTM network predicts the label (color) of the test data sequence, more precisely the label (color) of x_t. The output layer of the biLSTM network calculates a probability ($\hat{y}$) between 0 and 1. If $\hat{y}$ is greater than or equal to a threshold, which is set to 0.5, the biLSTM network outputs 1 and predicts x_t to be positive, i.e., predicts the label (color) of x_t to be blue; see Figure 1. Otherwise, the biLSTM network outputs 0 and predicts x_t to be negative, i.e., predicts the label (color) of x_t to be green; see Figure 1.

order. This design allows biLSTM to discover additional patterns that cannot be found by LSTM with only one recurrent layer (Siami-Namini et al. 2019). In addition, the data used in our study is time series and biLSTM has shown an improvement over LSTM for general time series forecasting (Althelaya et al. 2018; Kang et al. 2020). As our experimental results show later, biLSTM also outperforms LSTM in SEP prediction.

Figure 2 presents the architecture of our neural network, which accepts as input a data sequence with m consecutive data samples. (In the study presented here, m is set to 10.) The neural network consists of a biLSTM layer configured with 400 neurons. In addition, the neural network contains an attention layer motivated by Goodfellow et al. (2016) to direct the network to focus on important information and characteristics of input data samples. The attention layer is designed to map and capture the alignment between the input and output by calculating a weighted sum for input data sequences. Specifically, the attention context vector for the output $\hat{y}_i$, denoted $\boldsymbol{CV}_i$, is calculated as follows:

$$\boldsymbol{CV}_i = \sum_{j=1}^{m} \boldsymbol{W}_{i,j} \boldsymbol{H}_j, \tag{2}$$

where m is the input sequence length, $\boldsymbol{H}_j$ is the hidden state corresponding to the input data sample x_j and $\boldsymbol{W}$ contains weights applied to the hidden state. $\boldsymbol{W}$ is computed by a softmax function as follows:

$$\boldsymbol{W}_{i,j} = \frac{e^{S_{i,j}}}{\sum_{k=1}^{m} e^{S_{i,k}}}. \tag{3}$$

Here $S_{i,j}$ is a score function calculated as follows:

$$S_{i,j} = \boldsymbol{V} \times \tanh(\boldsymbol{W}^{'}(\boldsymbol{S}_i, \boldsymbol{H}_j)), \tag{4}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, $\boldsymbol{S}_i$ is the output state corresponding to the output $\hat{y}_i$, $\boldsymbol{V}$ and $\boldsymbol{W}^{'}$ are weight matrices learned by the neural network. The attention layer passes its resulting vector to a fully connected layer.

During training, our biLSTM network takes as input overlapping data sequences where each data sequence contains $m = 10$ consecutive data samples. The label (color) of a training data sequence is defined to be the label (color) of the

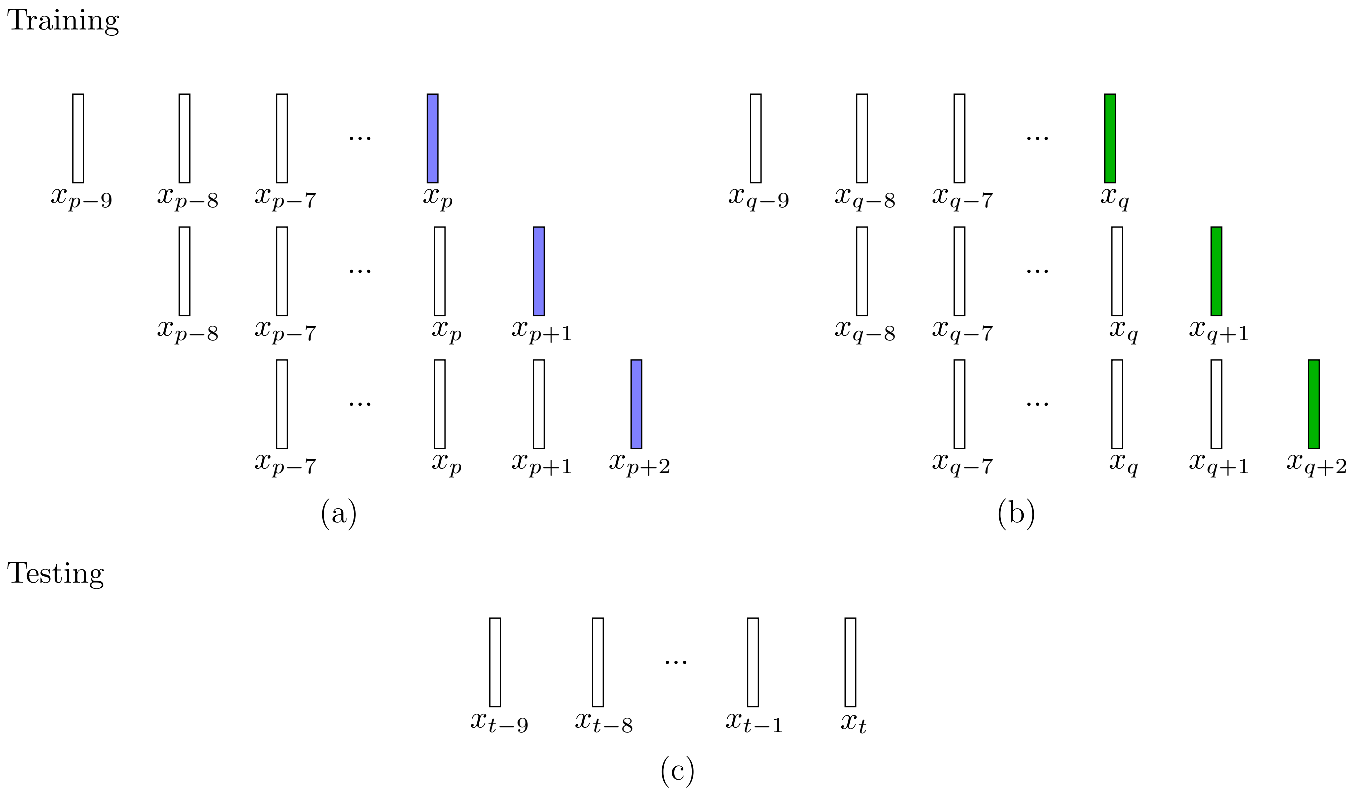


Figure 3. Example data sequences used to train and test our biLSTM network where each data sequence contains 10 consecutive data samples. (a) Three positive training data sequences taken from a flaring AR. (b) Three negative training data sequences taken from a flaring AR. In (a) and (b), the label (color) of a training data sequence is defined to be the label (color) of the last data sample in the training data sequence while the labels (colors) of the other nine data samples in the training data sequence are ignored. (c) A test data sequence formed for predicting the label (color) of the last data sample x_t in a flaring AR.

last (i.e., 10th) data sample in the data sequence while the labels (colors) of the other nine data samples in the data sequence are ignored. Thus, if the 10th data sample is positive (blue), then the training data sequence is positive; if the 10th data sample is negative (green), then the training data sequence is negative. We feed one training data sequence at a time to our biLSTM network when training the model. Figure 3(a) illustrates three positive data sequences used to train our biLSTM model. Figure 3(b) illustrates three negative data sequences used to train our biLSTM model.

The loss function used in our biLSTM model is the weighted binary cross-entropy (WBCE) (Goodfellow et al. 2016; Liu et al. 2020). Let N denote the total number of data sequences each having m consecutive data samples in the training set. Let w_0 denote the weight for the positive class (i.e., minority class) and let w_1 denote the weight for the negative class (i.e., majority class). The weights are calculated based on the ratio of majority and minority class sizes with more weight assigned to the minority class. Let y_i denote the observed probability of the ith data sequence; y_i is 1 if the ith data sequence is positive and 0 if the ith data sequence is negative. Let $\hat{y}_i$ denote the predicted probability of the ith data sequence. The WBCE, calculated as follows, is suitable for imbalanced datasets such as those tackled here where the negative class has more data samples than the positive class; see Table 1.

$$\mathrm{WBCE} = \sum_{i=1}^{N} w_0 y_i \log(\hat{y}_i) + w_1(1 - y_i)\log(1 - \hat{y}_i). \quad (5)$$

We configure the network to use a fraction (1/10) of the training set as the internal validation subset. We employ progressive learning with early stopping and adopt the strategy of saving the highest performing model during the iterative learning process. The performance of a model is measured by the WBCE on the internal validation subset where the smaller the WBCE is, the better performance the model has. In each iteration, the process checks the performance of the models in the current and previous iterations to decide which model to use for the next iteration. If the model in the current iteration has better performance, the process copies its weights as starting weights for the next iteration; otherwise, it copies the weights of the model in the previous iteration as starting weights for the next iteration. This progressive process improves the weights of the network's hidden layers and as a result the overall performance of the network is also improved. In addition, during the iterations, if the performance of the network degrades, the process stops and selects the highest performing model it identifies within the iterations.

During testing/prediction, we are given a test data sample x_t and our biLSTM model will predict the label (color) of x_t, i.e., predict whether x_t is positive or negative. We pack the $m-1$ data samples preceding x_t, namely x_{t-m+1}, x_{t-m+2}, ..., x_{t-1}, along with x_t into a test data sequence with m consecutive data samples and feed this test data sequence to our biLSTM model as shown in the input layer in Figure 2. Figure 3(c) illustrates a test data sequence where m is 10. The output layer of our biLSTM model calculates a probability between 0 and 1 for the test data sequence. We compare the probability with a threshold, which is set to 0.5. If the probability is greater than or equal to the threshold, our biLSTM model outputs 1 indicating the test data sequence, more precisely the test data sample x_t, is positive; otherwise our model outputs 0 indicating the test data sequence, more precisely x_t, is negative.

4. RESULTS

4.1. *Performance Metrics and Experiment Setup*

We conducted a series of experiments to evaluate the performance of the proposed method and compare it with related machine learning methods. For the data sample x_t at time point t, we define:

- x_t to be true positive (TP) if our model (network) predicts that x_t is positive and x_t is indeed positive, i.e., an SEP event will be produced with respect to x_t;
- x_t to be false positive (FP) if our model predicts that x_t is positive while x_t is actually negative, i.e., no SEP event will be produced with respect to x_t;
- x_t to be true negative (TN) if our model predicts x_t is negative and x_t is indeed negative;
- x_t to be false negative (FN) if our model predicts x_t is negative while x_t is actually positive.

We also use TP (FP, TN, FN, respectively) to denote the total number of true positives (false positives, true negatives, false negatives, respectively) produced by a method.

The following performance metrics are used in our study:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

$$\text{Balanced Accuracy (BACC)} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}\right), \tag{8}$$

$$\text{Heidke Skill Score (HSS)} = \frac{2 \times (\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{(\text{TP} + \text{FN}) \times (\text{FN} + \text{TN}) + (\text{TP} + \text{FP}) \times (\text{FP} + \text{TN})}, \tag{9}$$

$$\text{True Skill Statistics (TSS)} = \frac{\text{TP}}{\text{TP} + \text{FN}} - \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{10}$$

BACC (García et al. 2009) is an accuracy measure mainly for imbalanced datasets. HSS (Heidke 1926) and TSS (Bloomfield et al. 2012) are commonly used for flare, CME and SEP predictions (Bloomfield et al. 2012; Florios et al. 2018; Inceoglu et al. 2018; Liu et al. 2019; Liu et al. 2020). HSS ranges from $-\infty$ to $+1$. The higher HSS value a method has, the better performance the method achieves. TSS ranges from -1 to $+1$. Like HSS, the higher TSS value a method has, the better performance the method achieves. In addition, we use the weighted area under the curve (WAUC) (Bekkar et al. 2013) in our study. The area under the curve (AUC) in a receiver operating characteristic (ROC) curve (Marzban 2004) indicates how well a method is capable of distinguishing between two classes in binary prediction with the ideal value of one. When calculating the AUC, we do not distinguish between the accuracy on the minority class (positive class) and the accuracy on the majority class (negative class). In contrast, when calculating

Table 2. Importance Rankings of the 18 SHARP Parameters Used in Our Study for the FC_S and F_S Problems Respectively

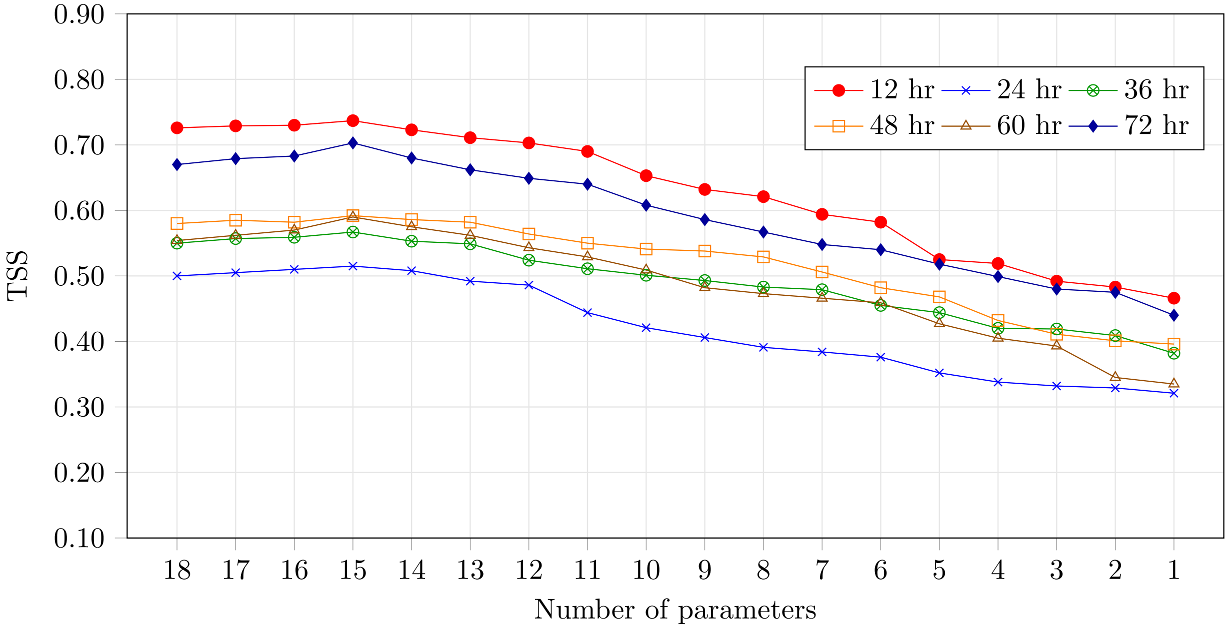
SHARP	12 hr		24 hr		36 hr		48 hr		60 hr		72 hr	
Keyword	FC_S	F_S	FC_S	F_S	FC_S	F_S	FC_S	F_S	FC_S	F_S	FC_S	F_S
ABSNJZH	3	3	1	4	1	10	1	10	1	2	5	1
AREA_ACR	17	16	17	16	17	16	17	16	16	16	16	16
MEANALP	13	15	3	15	3	15	3	15	2	15	6	15
MEANGAM	4	14	4	14	4	14	4	14	4	14	8	14
MEANGBH	5	13	5	13	5	13	5	13	14	13	14	7
MEANGBT	6	12	6	12	6	12	6	12	13	12	13	13
MEANGBZ	7	11	7	3	7	4	7	4	12	11	4	4
MEANJZD	8	10	8	11	8	11	8	11	11	10	12	12
MEANJZH	2	9	2	10	2	5	2	5	10	9	11	10
MEANPOT	10	8	10	9	10	9	10	9	9	8	1	11
MEANSHR	11	7	11	8	11	8	11	8	8	7	10	3
R_VALUE	12	6	12	7	12	1	12	3	7	6	2	2
SAVNCPP	1	2	13	2	13	3	13	1	6	1	3	5
SHRGT45	14	5	14	6	14	7	14	7	5	5	9	9
TOTPOT	15	4	15	5	15	6	15	6	15	4	15	8
TOTUSJH	9	1	9	1	9	2	9	2	3	3	7	6
TOTUSJZ	16	17	16	17	16	17	16	17	17	17	17	17
USFLUX	18	18	18	18	18	18	18	18	18	18	18	18

the WAUC, which is an extension of the AUC and mainly for imbalanced datasets like those tackled here, the accuracy on the minority class has a larger contribution to the overall performance of a model than the accuracy on the majority class. As a consequence, we assign more weight to the accuracy on the minority class where the weight is defined to be the ratio of the sizes of the minority and majority classes. All the metrics mentioned above are calculated using the confusion matrices obtained from the cross-validation (CV) scheme. With CV, we train a model using a subset of data, called the training set, and test the model using another subset of data, called the test set, where the training set and test set are disjoint. We consider six years, namely 2011-2015 and 2017, as mentioned in Section 2. Data samples from each year in turn are used for testing in a run and data samples from all the other five years together are used for training in the run. There are six years, and hence there are six runs in total. For each performance metric, the mean and standard deviation over the six runs are calculated and recorded.

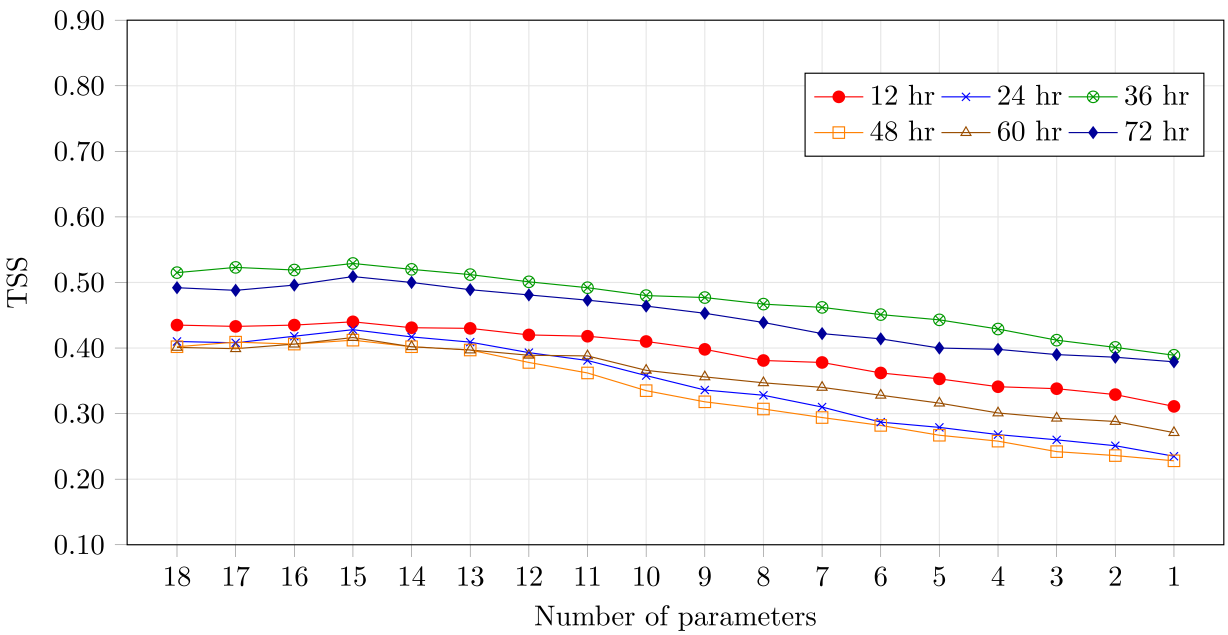
4.2. *Parameter Ranking and Selection*

We first assessed the importance of the 18 SHARP parameters described in Section 2 to understand which parameters are the most important ones with the greatest predictive power by utilizing a parameter ranking method, called Stability Selection (Meinshausen & Bühlmann 2010). This method is based on the LASSO (Least Absolute Shrinkage and Selection Operator) algorithm (Tibshirani 1996). Table 2 presents the rankings of the parameters with respect to T = 12, 24, 36, 48, 60 and 72 for the FC_S and F_S problems respectively. The parameter ranked first is the most important one while the parameter ranked 18th is the least important one. ABSNJZH is ranked consistently high for the FC_S problem while SAVNCPP and TOTUSJH are ranked high for the F_S problem. AREA_ACR, TOTUSJZ and USFLUX are ranked consistently low for both of the FC_S and F_S problems.

We then used the recursive parameter elimination algorithm (Butcher & Smith 2020) in combination with our biLSTM model to select a set of parameters that achieves the best performance where the performance is measured by TSS. The parameter elimination algorithm is an interactive procedure. It selects parameters by recursively considering smaller and smaller sets of parameters where the least important parameters are successively pruned from the current set of parameters. Figure 4 presents the parameter selection results for the FC_S and F_S problems respectively. It



(a)



(b)

Figure 4. Parameter selection results for (a) the FC_S problem, and (b) the F_S problem.

can be seen from the figure that using the top 15 most important parameters achieves the best performance for both of the FC_S and F_S problems. When using the top k, $1 \leq k \leq 14$, most important parameters, the less parameters we use, the worse performance our model achieves. Using the top-ranked, most important parameter alone would yield a lower TSS than using all the top 15 most important parameters together. In subsequent experiments, we used the top 15 most important parameters for our biLSTM model. That is, we removed the three least important parameters AREA_ACR, TOTUSJZ and USFLUX from data samples and each data sample contained only the top 15 most important SHARP parameters.

4.3. *Performance Comparison*

Next, we compared our biLSTM network with four related machine learning methods, including multilayer perceptrons (MLP), support vector machines (SVM), random forests (RF) and long short-term memory (LSTM) (Liu et al. 2019). These four methods are commonly used to predict solar flares, CMEs and SEPs (Bobra & Ilonidis 2016; Liu et al. 2017; Florios et al. 2018; Inceoglu et al. 2018; Chen et al. 2019; Liu et al. 2019; Liu et al. 2020; Wang et al. 2020; Abduallah et al. 2021).

MLP (Rosenblatt 1958; Arias del Campo et al. 2021) is a feed-forward artificial neural network (Braspenning et al. 1995) that consists of an input layer, an output layer, and one or more hidden layers. The number of hidden layers is set to 3 with 200 neurons in each hidden layer. SVM (Cristianini & Ricci 2008) is trained with the Radial Basis Function (RBF) kernel and the cache size is set to 20000 to speed the training process. RF (Breiman et al. 1984) is an ensemble algorithm that has two hyperparameters for performance tuning: m (the number of SHARP magnetic parameters randomly selected and used to split a node in a tree of the forest) and n (the number of trees to grow). We set m to 2 and n to 500. The implementation of LSTM follows that described in Liu et al. (2020). The hyperparameters not specified here are set to their default values provided by the scikit-learn library in Python (Pedregosa et al. 2011).

As done in biLSTM, we used the recursive parameter elimination algorithm (Butcher & Smith 2020) to identify and select the best parameters for the four related machine learning methods based on the importance rankings of the 18 SHARP parameters in Table 2 for the FC_S and F_S problems respectively. Our experiments showed that, like biLSTM, using the top 15 most important parameters achieved the best performance for the four related machine learning methods. Consequently, we used the top 15 most important parameters for the four machine learning methods in the experimental study.

Figures 5 and 6 present the confusion matrices of the five machine learning methods (RF, MLP, SVM, LSTM, biLSTM) for the FC_S and F_S problems respectively. For each T, T = 12, 24, 36, 48, 60, 72, and each machine learning method, the figures show the minimum, average, maximum (displayed from top to bottom) TP, FN, TN, FP respectively from the six runs based on our cross-validation scheme. For example, refer to $T = 12$ and biLSTM in Figure 5. The minimum (maximum, respectively) TP obtained by biLSTM from the six runs is 87 (246, respectively); the average TP over the six runs is 139. It can be seen from Figures 5 and 6 that the average TN values are much larger than the average FP values for both of the FC_S and F_S problems. This happens because there are many negative training data samples in our datasets (see Table 1). As a consequence, the machine learning methods gain sufficient knowledge about the negative data samples and hence can detect them relatively easily. For the FC_S problem, the average TP values (TN values, respectively) are consistently larger than the average FN values (FP values, respectively), indicating that the machine learning methods can solve the FC_S problem reasonably well. For the F_S problem, the average TP values are close to, or even smaller than, the average FN values in many cases, suggesting that the machine learning methods have difficulty in detecting positive data samples. This is understandable given that there are much fewer positive training data samples than negative training data samples for the F_S problem (see Table 1).

Tables 3 and 4 compare the performance of the five machine learning methods for the FC_S and F_S problems respectively. The tables present the mean performance metric values averaged over the six runs based on our cross-validation scheme with standard deviations enclosed in parentheses. Best average metric values are highlighted in boldface. It can be seen from Tables 3 and 4 that our biLSTM network outperforms the four related machine learning methods in terms of BACC, HSS, TSS and WAUC. Furthermore, the five machine learning methods generally perform better in solving the FC_S problem than in solving the F_S problem. This result indicates that one can predict SEP events more accurately when active regions will produce both flares and associated CMEs. Using flare information alone to predict SEP events is harder and would produce less reliable prediction results.

4.4. *Probabilistic Forecasting and Calibration*

The five machine learning methods (RF, MLP, SVM, LSTM, biLSTM) studied here are inherently probabilistic forecasting models in the sense that they calculate a probability between 0 and 1. We compare the probability with a threshold, which is set to 0.5, to determine the output produced by each machine learning method. The output is either 1 or 0 (see Figure 2), and hence each method is essentially a binary prediction model. In addition to comparing the methods used as binary prediction models, we also compare the methods used as probabilistic forecasting models, where the output produced by each model is interpreted as follows. **[FC_S problem]** Given a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t and the flare initiates a CME, based on the SHARP parameters in x_t and its preceding $m-1$ data samples x_{t-m+1}, x_{t-m+2}, …, x_{t-1}, we

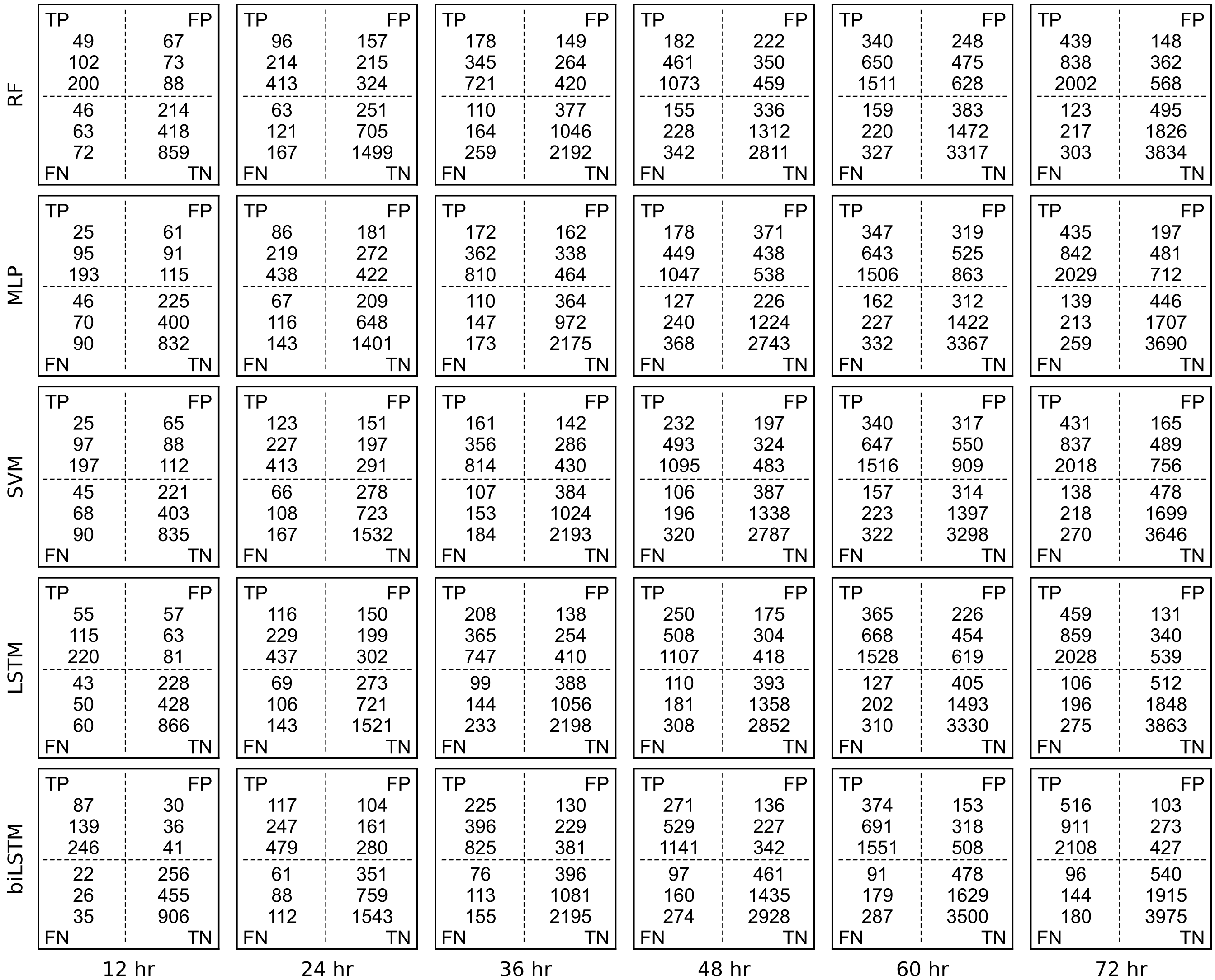


Figure 5. Confusion matrices of RF, MLP, SVM, LSTM and biLSTM for the FC_S problem. For each T, T = 12, 24, 36, 48, 60, 72, and each machine learning method, the figure shows the minimum, average, maximum (displayed from top to bottom) TP, FN, TN, FP respectively from the six runs based on our cross-validation scheme.

calculate and output a probabilistic estimate of how likely it is that the AR will produce an SEP event associated with the flare and CME. [**F_S problem**] Given a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t regardless of whether or not the flare initiates a CME, based on the SHARP parameters in x_t and its preceding $m-1$ data samples x_{t-m+1}, x_{t-m+2}, ..., x_{t-1}, we calculate and output a probabilistic estimate of how likely it is that the AR will produce an SEP event associated with the flare.

The distribution and behavior of the predicted probabilistic values may not match the expected distribution of observed probabilities in the training data. One can adjust the distribution of the predicted probabilities to better match the expected distribution observed in the training data through calibration. Here, we adopt isotonic regression (Kruskal 1964; Sager & Thisted 1982) to adjust the probabilities. Isotonic regression works by fitting a free-form line to a sequence of data points such that the fitted line is non-decreasing (or non-increasing) everywhere, and lies as close to the data points as possible. Calibrated models often produce more accurate results. We add a suffix "+C" to each model to denote the calibrated version of the model.

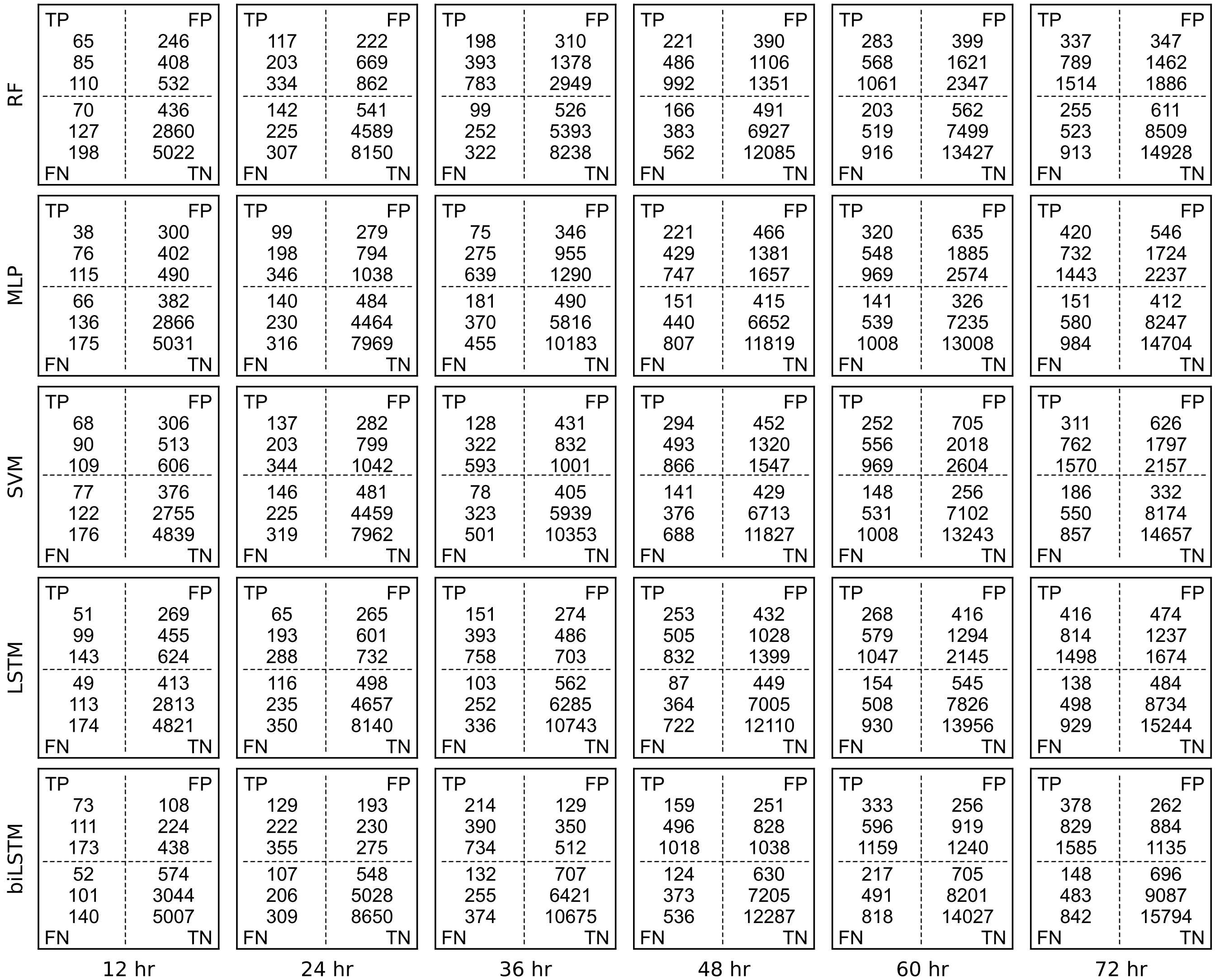


Figure 6. Confusion matrices of RF, MLP, SVM, LSTM and biLSTM for the F_S problem. For each T, T = 12, 24, 36, 48, 60, 72, and each machine learning method, the figure shows the minimum, average, maximum (displayed from top to bottom) TP, FN, TN, FP respectively from the six runs based on our cross-validation scheme.

To quantitatively assess the performance of a probabilistic forecasting model, we adopt the Brier Score (BS; Wilks 2010) and Brier Skill Score (BSS; Wilks 2010), defined as follows:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{11}$$

$$\mathrm{BSS} = 1 - \frac{BS}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{12}$$

Here, N is the total number of data sequences each having m consecutive data samples in the test set (see Figure 2 where a test data sequence with m consecutive data samples is fed to our biLSTM model); y_i denotes the observed probability and $\hat{y}_i$ denotes the predicted probability of the ith test data sequence respectively; $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ denotes the mean of all the observed probabilities. The BS values range from 0 to 1, with 0 being a perfect score, whereas the BSS values range from $-\infty$ to 1, with 1 being a perfect score.

Table 5 compares the performance of the five machine learning methods used as probabilistic forecasting models for the FC_S and F_S problems respectively. The table presents the mean BS and BSS values averaged over the six runs

Table 3. Performance Comparison of RF, MLP, SVM, LSTM and biLSTM Based on Our Cross-Validation Scheme for the FC_S Problem

		12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
Recall	RF	0.593 (0.106)	0.617 (0.120)	0.660 (0.078)	0.632 (0.134)	0.721 (0.082)	0.767 (0.095)
	MLP	0.542 (0.175)	0.617 (0.149)	0.679 (0.110)	0.618 (0.143)	0.711 (0.088)	0.766 (0.093)
	SVM	0.543 (0.175)	0.666 (0.088)	0.663 (0.125)	0.686 (0.123)	0.714 (0.090)	0.761 (0.093)
	LSTM	0.658 (0.113)	0.663 (0.099)	0.699 (0.064)	0.714 (0.100)	0.745 (0.071)	0.788 (0.089)
	biLSTM	**0.825** (0.058)	**0.710** (0.112)	**0.758** (0.064)	**0.748** (0.086)	**0.777** (0.077)	**0.843** (0.058)
Precision	RF	0.554 (0.113)	0.483 (0.151)	0.556 (0.163)	0.532 (0.165)	0.548 (0.155)	0.661 (0.171)
	MLP	0.478 (0.153)	0.432 (0.166)	0.501 (0.155)	0.471 (0.161)	0.524 (0.164)	0.605 (0.161)
	SVM	0.486 (0.168)	0.520 (0.132)	0.543 (0.158)	0.574 (0.163)	0.514 (0.176)	0.603 (0.175)
	LSTM	0.609 (0.102)	0.516 (0.142)	0.581 (0.158)	0.599 (0.155)	0.569 (0.153)	0.680 (0.166)
	biLSTM	**0.777** (0.061)	**0.587** (0.148)	**0.619** (0.149)	**0.669** (0.155)	**0.656** (0.160)	**0.739** (0.133)
BACC	RF	0.707 (0.045)	0.672 (0.047)	0.719 (0.036)	0.688 (0.051)	0.713 (0.030)	0.790 (0.043)
	MLP	0.663 (0.061)	0.641 (0.062)	0.690 (0.049)	0.641 (0.051)	0.691 (0.046)	0.752 (0.023)
	SVM	0.668 (0.070)	0.706 (0.022)	0.709 (0.045)	0.730 (0.032)	0.687 (0.053)	0.753 (0.029)
	LSTM	0.751 (0.047)	0.704 (0.035)	0.744 (0.029)	0.750 (0.031)	0.734 (0.025)	0.807 (0.041)
	biLSTM	**0.868** (0.029)	**0.757** (0.040)	**0.784** (0.029)	**0.796** (0.037)	**0.795** (0.032)	**0.852** (0.021)
HSS	RF	0.404 (0.091)	0.315 (0.100)	0.406 (0.088)	0.354 (0.100)	0.386 (0.063)	0.545 (0.099)
	MLP	0.314 (0.119)	0.249 (0.123)	0.343 (0.105)	0.259 (0.098)	0.343 (0.098)	0.468 (0.067)
	SVM	0.325 (0.138)	0.377 (0.063)	0.391 (0.101)	0.431 (0.080)	0.332 (0.114)	0.468 (0.083)
	LSTM	0.489 (0.090)	0.373 (0.086)	0.450 (0.084)	0.468 (0.079)	0.423 (0.057)	0.579 (0.099)
	biLSTM	**0.722** (0.057)	**0.481** (0.097)	**0.522** (0.080)	**0.562** (0.086)	**0.551** (0.086)	**0.669** (0.064)
TSS	RF	0.413 (0.090)	0.344 (0.094)	0.437 (0.072)	0.376 (0.101)	0.426 (0.061)	0.579 (0.085)
	MLP	0.326 (0.123)	0.281 (0.125)	0.379 (0.098)	0.283 (0.101)	0.382 (0.092)	0.504 (0.046)
	SVM	0.336 (0.140)	0.413 (0.045)	0.417 (0.091)	0.459 (0.063)	0.374 (0.106)	0.507 (0.057)
	LSTM	0.501 (0.093)	0.407 (0.071)	0.487 (0.059)	0.499 (0.063)	0.468 (0.051)	0.615 (0.082)
	biLSTM	**0.737** (0.057)	**0.515** (0.081)	**0.567** (0.059)	**0.592** (0.073)	**0.590** (0.063)	**0.703** (0.041)
WAUC	RF	0.453 (0.056)	0.375 (0.071)	0.476 (0.048)	0.410 (0.063)	0.459 (0.040)	0.621 (0.057)
	MLP	0.354 (0.033)	0.301 (0.032)	0.415 (0.088)	0.304 (0.072)	0.410 (0.024)	0.543 (0.085)
	SVM	0.361 (0.087)	0.453 (0.068)	0.457 (0.023)	0.503 (0.064)	0.405 (0.040)	0.553 (0.085)
	LSTM	0.541 (0.074)	0.436 (0.071)	0.526 (0.020)	0.535 (0.052)	0.510 (0.051)	0.671 (0.026)
	biLSTM	**0.794** (0.041)	**0.563** (0.039)	**0.609** (0.086)	**0.646** (0.043)	**0.642** (0.048)	**0.764** (0.073)

based on our cross-validation scheme with standard deviations enclosed in parentheses. Best BS and BSS values are highlighted in boldface. It can be seen from Table 5 that the probabilistic forecasting models generally perform better in solving the FC_S problem than in solving the F_S problem, suggesting that F_S is a harder problem and hence the forecasting results for the F_S problem would be less reliable. These findings are consistent with those in Tables 3 and 4 where the machine learning methods are used as binary prediction models. Furthermore, the calibrated version of a model is better than the model without calibration. Overall, biLSTM+C performs the best among all the models in terms of both BS and BSS.

5. DISCUSSION AND CONCLUSIONS

We develop a bidirectional long short-term memory (biLSTM) network for SEP prediction. We consider two prediction tasks. In the first task (FC_S), given a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t and the flare initiates a CME, based on the SHARP parameters in x_t and its preceding $m-1$ data samples x_{t-m+1}, x_{t-m+2}, …, x_{t-1}, our biLSTM, when used as a binary prediction model, can predict whether the AR will produce an SEP event associated with the flare/CME. Furthermore, our biLSTM,

Table 4. Performance Comparison of RF, MLP, SVM, LSTM and biLSTM Based on Our Cross-Validation Scheme for the F_S Problem

		12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
Recall	RF	0.414 (0.099)	0.468 (0.066)	0.592 (0.137)	0.550 (0.123)	0.522 (0.120)	0.590 (0.109)
	MLP	0.367 (0.138)	0.457 (0.098)	0.398 (0.172)	0.501 (0.148)	0.520 (0.153)	0.568 (0.150)
	SVM	0.433 (0.064)	0.471 (0.058)	0.495 (0.189)	0.575 (0.106)	0.518 (0.160)	0.567 (0.145)
	LSTM	0.468 (0.146)	0.456 (0.166)	0.592 (0.155)	**0.591** (0.140)	0.546 (0.155)	0.624 (0.127)
	biLSTM	**0.520** (0.103)	**0.508** (0.122)	**0.597** (0.095)	0.545 (0.197)	**0.548** (0.090)	**0.629** (0.120)
Precision	RF	0.178 (0.052)	0.253 (0.119)	0.267 (0.159)	0.314 (0.144)	0.285 (0.166)	0.370 (0.177)
	MLP	0.164 (0.073)	0.215 (0.106)	0.227 (0.140)	0.254 (0.136)	0.247 (0.138)	0.315 (0.158)
	SVM	0.152 (0.035)	0.218 (0.096)	0.278 (0.130)	0.287 (0.129)	0.234 (0.136)	0.306 (0.154)
	LSTM	0.184 (0.076)	0.252 (0.114)	0.432 (0.124)	0.340 (0.141)	0.329 (0.162)	0.404 (0.161)
	biLSTM	**0.366** (0.159)	**0.473** (0.110)	**0.527** (0.155)	**0.377** (0.200)	**0.405** (0.170)	**0.485** (0.166)
BACC	RF	0.627 (0.033)	0.656 (0.030)	0.684 (0.052)	0.681 (0.038)	0.650 (0.025)	0.702 (0.036)
	MLP	0.599 (0.049)	0.635 (0.031)	0.605 (0.062)	0.634 (0.036)	0.618 (0.040)	0.665 (0.032)
	SVM	0.616 (0.052)	0.641 (0.028)	0.655 (0.039)	0.676 (0.035)	0.605 (0.055)	0.654 (0.052)
	LSTM	0.647 (0.046)	0.653 (0.079)	0.739 (0.061)	0.703 (0.022)	0.677 (0.059)	0.720 (0.029)
	biLSTM	**0.721** (0.046)	**0.714** (0.047)	**0.765** (0.037)	**0.706** (0.076)	**0.708** (0.015)	**0.754** (0.027)
HSS	RF	0.154 (0.031)	0.218 (0.066)	0.239 (0.117)	0.265 (0.077)	0.212 (0.069)	0.311 (0.086)
	MLP	0.128 (0.073)	0.171 (0.049)	0.156 (0.098)	0.181 (0.050)	0.152 (0.048)	0.235 (0.049)
	SVM	0.119 (0.040)	0.175 (0.038)	0.233 (0.082)	0.235 (0.042)	0.128 (0.064)	0.218 (0.087)
	LSTM	0.170 (0.064)	0.216 (0.091)	0.414 (0.110)	0.303 (0.068)	0.270 (0.100)	0.352 (0.078)
	biLSTM	**0.365** (0.137)	**0.418** (0.100)	**0.493** (0.099)	**0.345** (0.159)	**0.353** (0.079)	**0.448** (0.089)
TSS	RF	0.254 (0.066)	0.313 (0.060)	0.368 (0.103)	0.362 (0.075)	0.301 (0.050)	0.405 (0.071)
	MLP	0.198 (0.098)	0.270 (0.062)	0.211 (0.125)	0.269 (0.072)	0.236 (0.080)	0.329 (0.064)
	SVM	0.233 (0.104)	0.282 (0.057)	0.309 (0.078)	0.353 (0.070)	0.210 (0.110)	0.309 (0.104)
	LSTM	0.293 (0.092)	0.305 (0.158)	0.479 (0.122)	0.405 (0.044)	0.353 (0.119)	0.440 (0.058)
	biLSTM	**0.441** (0.093)	**0.428** (0.093)	**0.529** (0.075)	**0.412** (0.151)	**0.416** (0.031)	**0.509** (0.055)
WAUC	RF	0.276 (0.022)	0.338 (0.067)	0.403 (0.021)	0.388 (0.035)	0.330 (0.021)	0.431 (0.042)
	MLP	0.212 (0.084)	0.290 (0.079)	0.226 (0.024)	0.294 (0.034)	0.256 (0.085)	0.359 (0.075)
	SVM	0.254 (0.084)	0.307 (0.048)	0.334 (0.062)	0.381 (0.074)	0.230 (0.060)	0.334 (0.027)
	LSTM	0.319 (0.056)	0.328 (0.047)	0.514 (0.033)	0.432 (0.038)	0.382 (0.073)	0.474 (0.020)
	biLSTM	**0.480** (0.076)	**0.467** (0.034)	**0.574** (0.085)	**0.450** (0.035)	**0.448** (0.087)	**0.552** (0.016)

when used as a probabilistic forecasting model, can provide a probabilistic estimate of how likely it is that the AR will produce an SEP event associated with the flare/CME. In the second task (F_S), given a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t, based on the SHARP parameters in x_t and its preceding $m-1$ data samples x_{t-m+1}, x_{t-m+2}, ..., x_{t-1}, our biLSTM, when used as a binary prediction model, can predict whether the AR will produce an SEP event associated with the flare, and when used as a probabilistic forecasting model, can provide a probabilistic estimate of how likely it is that the AR will produce an SEP event associated with the flare, regardless of whether or not the flare initiates a CME. For both tasks, T ranges from 12 to 72 in 12 hr intervals.

We surveyed and collected data samples from the JSOC website, in the period between 2010 and 2021. Each data sample contains 18 SHARP parameters. Active regions (ARs) from 2010, 2016, and 2018-2021 were excluded from the study due to the lack of qualified data samples or the absence of SEP events associated with M-/X-class flares and CMEs. We then performed a cross-validation study on the remaining six years (2011-2015 and 2017). In the cross-validation study, training and test sets are disjoint, and hence our biLSTM model can make predictions on ARs

Table 5. Probabilistic Forecasting Results of RF, MLP, SVM, LSTM and biLSTM With and Without Calibration for the FC_S and F_S Problems Respectively

			12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
FC_S	BS	RF	0.372 (0.083)	0.342 (0.075)	0.365 (0.060)	0.342 (0.092)	0.355 (0.051)	0.269 (0.040)
		RF+C	0.332 (0.070)	0.331 (0.101)	0.324 (0.056)	0.302 (0.085)	0.328 (0.047)	0.252 (0.037)
		MLP	0.362 (0.136)	0.393 (0.080)	0.335 (0.095)	0.315 (0.050)	0.335 (0.083)	0.280 (0.025)
		MLP+C	0.329 (0.124)	0.366 (0.076)	0.309 (0.088)	0.283 (0.050)	0.301 (0.073)	0.255 (0.027)
		SVM	0.359 (0.052)	0.344 (0.037)	0.337 (0.075)	0.353 (0.049)	0.306 (0.087)	0.298 (0.034)
		SVM+C	0.322 (0.047)	0.303 (0.030)	0.297 (0.066)	0.306 (0.035)	0.284 (0.080)	0.267 (0.035)
		LSTM	0.337 (0.064)	0.356 (0.054)	0.300 (0.058)	0.288 (0.032)	0.298 (0.038)	0.274 (0.028)
		LSTM+C	0.271 (0.062)	0.302 (0.045)	0.262 (0.053)	0.244 (0.032)	0.273 (0.035)	0.232 (0.021)
		biLSTM	0.248 (0.028)	0.272 (0.052)	0.270 (0.048)	0.281 (0.040)	0.297 (0.021)	0.279 (0.015)
		biLSTM+C	**0.215** (0.021)	**0.249** (0.038)	**0.223** (0.025)	**0.235** (0.033)	**0.270** (0.043)	**0.202** (0.014)
	BSS	RF	0.273 (0.166)	0.316 (0.164)	0.282 (0.124)	0.316 (0.180)	0.325 (0.118)	0.466 (0.067)
		RF+C	0.341 (0.140)	0.343 (0.189)	0.362 (0.115)	0.396 (0.167)	0.340 (0.109)	0.501 (0.062)
		MLP	0.290 (0.262)	0.274 (0.100)	0.320 (0.202)	0.382 (0.082)	0.323 (0.179)	0.436 (0.048)
		MLP+C	0.355 (0.239)	0.325 (0.094)	0.372 (0.187)	0.445 (0.087)	0.392 (0.158)	0.486 (0.052)
		SVM	0.281 (0.128)	0.310 (0.086)	0.333 (0.142)	0.295 (0.101)	0.388 (0.178)	0.406 (0.057)
		SVM+C	0.355 (0.115)	0.392 (0.065)	0.412 (0.125)	0.389 (0.074)	0.432 (0.164)	0.469 (0.059)
		LSTM	0.338 (0.124)	0.306 (0.114)	0.388 (0.128)	0.425 (0.073)	0.406 (0.074)	0.458 (0.052)
		LSTM+C	0.466 (0.121)	0.395 (0.099)	0.466 (0.115)	0.513 (0.068)	0.456 (0.068)	0.542 (0.037)
		biLSTM	0.513 (0.050)	0.450 (0.110)	0.464 (0.097)	0.424 (0.086)	0.417 (0.046)	0.450 (0.018)
		biLSTM+C	**0.578** (0.035)	**0.498** (0.080)	**0.558** (0.046)	**0.518** (0.065)	**0.470** (0.087)	**0.587** (0.020)
F_S	BS	RF	0.393 (0.094)	0.391 (0.075)	0.449 (0.126)	0.459 (0.077)	0.381 (0.063)	0.317 (0.056)
		RF+C	0.341 (0.078)	0.351 (0.062)	0.380 (0.109)	0.383 (0.063)	0.334 (0.043)	0.276 (0.044)
		MLP	0.433 (0.042)	0.429 (0.053)	0.395 (0.063)	0.404 (0.105)	0.394 (0.134)	0.366 (0.072)
		MLP+C	0.376 (0.031)	0.381 (0.046)	0.340 (0.057)	0.357 (0.100)	0.341 (0.111)	0.329 (0.069)
		SVM	0.429 (0.071)	0.391 (0.043)	0.381 (0.032)	0.379 (0.075)	0.403 (0.067)	0.390 (0.100)
		SVM+C	0.390 (0.073)	0.354 (0.038)	0.336 (0.025)	0.336 (0.061)	0.363 (0.067)	0.346 (0.082)
		LSTM	0.373 (0.093)	0.359 (0.077)	0.381 (0.048)	0.347 (0.042)	0.377 (0.074)	0.276 (0.034)
		LSTM+C	0.341 (0.088)	0.315 (0.074)	0.336 (0.049)	0.314 (0.036)	0.319 (0.052)	0.247 (0.034)
		biLSTM	0.267 (0.054)	0.345 (0.038)	0.318 (0.069)	0.344 (0.040)	0.346 (0.034)	0.307 (0.028)
		biLSTM+C	**0.231** (0.042)	**0.294** (0.035)	**0.226** (0.060)	**0.291** (0.044)	**0.289** (0.021)	**0.220** (0.029)
	BSS	RF	0.267 (0.182)	0.206 (0.160)	0.232 (0.224)	0.228 (0.066)	0.313 (0.131)	0.360 (0.093)
		RF+C	0.322 (0.150)	0.329 (0.133)	0.293 (0.216)	0.354 (0.078)	0.322 (0.093)	0.441 (0.072)
		MLP	0.282 (0.078)	0.138 (0.109)	0.226 (0.105)	0.284 (0.076)	0.259 (0.222)	0.345 (0.085)
		MLP+C	0.336 (0.059)	0.235 (0.093)	0.335 (0.088)	0.368 (0.069)	0.358 (0.183)	0.411 (0.086)
		SVM	0.122 (0.163)	0.228 (0.083)	0.251 (0.051)	0.247 (0.151)	0.204 (0.137)	0.310 (0.119)
		SVM+C	0.201 (0.165)	0.301 (0.075)	0.339 (0.043)	0.332 (0.122)	0.283 (0.135)	0.388 (0.091)
		LSTM	0.256 (0.204)	0.297 (0.150)	0.230 (0.092)	0.309 (0.080)	0.342 (0.152)	0.430 (0.129)
		LSTM+C	0.319 (0.193)	0.383 (0.144)	0.323 (0.096)	0.385 (0.067)	0.447 (0.145)	0.489 (0.122)
		biLSTM	0.464 (0.106)	0.309 (0.092)	0.451 (0.144)	0.314 (0.083)	0.351 (0.072)	0.437 (0.091)
		biLSTM+C	**0.535** (0.083)	**0.410** (0.084)	**0.513** (0.100)	**0.420** (0.087)	**0.457** (0.047)	**0.521** (0.089)

that were never seen before. We evaluated the performance of our model and compared it with four related machine learning algorithms, namely RF (Liu et al. 2017), MLP (Inceoglu et al. 2018), SVM (Bobra & Ilonidis 2016) and a previous LSTM network (Liu et al. 2019). The five machine learning methods including our biLSTM can be used both as binary prediction models and as probabilistic forecasting models. Our main results are summarized as follows.

1. The data samples in an AR are modeled as time series. We employ the biLSTM network to predict SEP events based on the time series. To our knowledge, this is the first study using a deep neural network to learn the dependencies in the temporal domain of the data for SEP prediction.

2. We evaluate the importance of the 18 SHARP parameters used in our study. It is found that using the top 15 SHARP parameters achieves the best performance for both the FC_S and F_S tasks. This finding is consistent with the literature which indicates using fewer high-quality SHARP parameters often achieves better performance for eruption prediction than using all the SHARP parameters including low-quality ones (Alpaydin 2016; Bobra & Ilonidis 2016; Liu et al. 2020).

3. Our experiments show that the proposed biLSTM outperforms the four related machine learning methods in performing binary prediction and probabilistic forecasting for both the FC_S and F_S tasks. Furthermore, we introduce a calibration mechanism to enhance the accuracy of probabilistic forecasting. Overall, the calibrated biLSTM achieves the best performance among all the probabilistic forecasting models studied here.

4. When both an M-/X-class flare and its associated CME will occur, predicting whether there is an SEP event associated with the flare and CME is an easier problem (FC_S). Our biLSTM can solve the FC_S problem with relatively high accuracy. In contrast, when an M-/X-class flare will occur in the absence of CME information, predicting whether there is an SEP event associated with the flare is a harder problem (F_S). Our biLSTM solves the F_S problem with relatively low accuracy, and hence the prediction results would be less reliable.

5. The findings reported here are based on the cross-validation (CV) scheme in which six years (2011-2015 and 2017) are considered, data samples from each year in turn are used for testing, and data samples from the other five years together are used for training. To further understand the behavior of our biLSTM network and the four related machine learning methods, we have performed additional experiments using a random division (RD) scheme. With RD, we randomly select 10% of all positive data sequences and 10% of all negative data sequences, and use them together as the test set. The remaining 90% of the positive data sequences and 90% of the negative data sequences are used together as the training set. We repeat this experiment 100 times. The average values and standard deviations of the performance metrics are calculated. Tables 6 and 7 in the Appendix present results of the five machine learning methods used as binary prediction models for the FC_S and F_S problems respectively. Table 8 presents results of the five machine learning methods used as probabilistic forecasting models for the FC_S and F_S problems respectively. It can be seen from these tables that the results obtained from the random division scheme are consistent with those from the cross-validation scheme, though the performance metric values from the RD scheme are generally better than those from the CV scheme. This happens probably because with the RD scheme the machine learning methods are trained by more diverse data and hence are more knowledgeable, yielding more accurate results than with the CV scheme.

It should be pointed out that, in solving the FC_S problem, the condition in which we have a data sample x_t at time point t in an AR where the AR will produce an M- or X-class flare within the next T hours of t and the flare initiates a CME is given. That is, we assume an M- or X-class flare and its associated CME will occur. In an operational system, one can determine in two phases if an AR will produce an M- or X-class flare within the next T-hours of a given time point t and if the flare initiates a CME, as follows (Liu 2020). In the first phase, one can use a flare prediction tool (e.g., Liu et al. 2017; Florios et al. 2018; Jonas et al. 2018; Nishizuka et al. 2018; Liu et al. 2019) to predict whether there will be an M- or X-class flare within the next T hours of t. If the answer is yes, then in the second phase one can use a CME prediction tool (e.g., Liu et al. 2020) to predict whether the flare initiates a CME. If the answer is also yes, then one can use the proposed biLSTM to predict whether there is an SEP event associated with the flare and CME. On the other hand, to solve the F_S problem, one only needs to execute the first phase. If the answer from the first phase indicates that an M- or X-class flare will occur within the next T hours of t, one can then go ahead to use the proposed biLSTM to predict whether there is an SEP event associated with the flare. Thus, the proposed biLSTM does not function in a stand-alone manner. Rather, it first requires the other tools to provide flare/CME predictions. As such, the performance of the operational biLSTM system depends on the performance of the other tools. A wrong prediction from the other tools would affect the accuracy of our approach.

We thank the referee and scientific editor for very helpful and thoughtful comments. We also thank the team of *SDO*/HMI for producing vector magnetic data products. The flare catalogs were prepared by and made available through NOAA NCEI. The CME and SEP event records were provided by DONKI. This work was supported by U.S. NSF grants AGS-1927578 and AGS-1954737. J.W. thanks Manolis K. Georgoulis for helpful conversations in the SHINE 2019 Conference. Q.L. and H.W. acknowledge the support of NASA under grants 80NSSC18K1705, 80NSSC19K0068 and 80NSSC20K1282.

APPENDIX

Tables 6 and 7 present results of the five machine learning methods (RF, MLP, SVM, LSTM, biLSTM) used as binary prediction models for the FC_S and F_S problems respectively. Table 8 presents results of the five machine learning methods used as probabilistic forecasting models for the FC_S and F_S problems respectively. The tables show the mean performance metric values averaged over the 100 experiments based on the random division scheme with standard deviations enclosed in parentheses. Best average metric values are highlighted in boldface.

REFERENCES

Abduallah, Y., Wang, J. T. L., Nie, Y., Liu, C., & Wang, H. 2021, Research in Astronomy and Astrophysics, 21, 160, doi: 10.1088/1674-4527/21/7/160

Ahmed, O. W., Qahwaji, R., Colak, T., et al. 2013, SoPh, 283, 157, doi: 10.1007/s11207-011-9896-1

Alpaydin, E. 2016, Machine Learning: The New AI, The MIT Press Essential Knowledge series (Cambridge, MA: MIT Press). https://mitpress.mit.edu/books/machine-learning

Althelaya, K. A., El-Alfy, E.-S. M., & Mohammed, S. 2018, in 2018 9th International Conference on Information and Communication Systems (ICICS), 151–156, doi: 10.1109/IACS.2018.8355458

Arias del Campo, F., Guevara Neri, M. C., Vergara Villegas, O. O., et al. 2021, Expert Systems with Applications, 181, 115147, doi: 10.1016/j.eswa.2021.115147

Bekkar, M., Djemaa, H. K., & Alitouche, T. A. 2013, Journal of Information Engineering and Applications, 3, 27. https://www.iiste.org/Journals/index.php/JIEA/article/view/7633/8051

Benz, A. O. 2008, Living Reviews in Solar Physics, 5, 1, doi: 10.12942/lrsp-2008-1

Berkebile-Stoiser, S., Veronig, A. M., Bein, B. M., & Temmer, M. 2012, ApJ, 753, 88, doi: 10.1088/0004-637X/753/1/88

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. 2012, ApJL, 747, L41, doi: 10.1088/2041-8205/747/2/L41

Bobra, M., Sun, X., Hoeksema, J., et al. 2014, SoPh, 289, 3549, doi: 10.1007/s11207-014-0529-3

Bobra, M. G., & Ilonidis, S. 2016, ApJ, 821, 127, doi: 10.3847/0004-637X/821/2/127

Braspenning, P. J., Thuijsman, F., & Weijters, A. J. M. M., eds. 1995, Lecture Notes in Computer Science, Vol. 931, Artificial Neural Networks: An Introduction to ANN Theory and Practice (Springer), doi: 10.1007/BFb0027019

Breiman, L., Friedman, J. H., & Olshen, R. A. 1984, Classification and Regression Trees (Boca Raton, Florida: Chapman and Hall/CRC), doi: 10.1201/9781315139470

Brito, T. V., Woodroffe, J., Jordanova, V. K., Henderson, M., & Birn, J. 2018, Journal of Atmospheric and Solar-Terrestrial Physics, 177, 131, doi: 10.1016/j.jastp.2017.10.008

Butcher, B., & Smith, B. J. 2020, The American Statistician, 74, 308, doi: 10.1080/00031305.2020.1790217

Chen, P. F. 2011, Living Reviews in Solar Physics, 8, 1, doi: 10.12942/lrsp-2011-1

Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, Space Weather, 17, 1404, doi: 10.1029/2019SW002214

Cristianini, N., & Ricci, E. 2008, Support Vector Machines (Boston, MA: Springer US), 928–932, doi: 10.1007/978-0-387-30162-4_415

Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, SoPh, 293, 28, doi: 10.1007/s11207-018-1250-4

García, V., Mollineda, R. A., & Sánchez, J. S. 2009, in Pattern Recognition and Image Analysis, ed. H. Araujo, A. M. Mendonça, A. J. Pinho, & M. I. Torres (Berlin, Heidelberg: Springer Berlin Heidelberg), 441–448, doi: 10.1007/978-3-642-02172-5_57

Table 6. Performance Comparison of RF, MLP, SVM, LSTM and biLSTM Based on the Random Division Scheme for the FC_S Problem

		12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
Recall	RF	0.699 (0.149)	0.686 (0.148)	0.657 (0.169)	0.701 (0.165)	0.689 (0.087)	0.817 (0.080)
	MLP	0.655 (0.164)	0.690 (0.132)	0.688 (0.149)	0.692 (0.149)	0.679 (0.101)	0.798 (0.091)
	SVM	0.666 (0.170)	0.693 (0.147)	0.681 (0.159)	0.701 (0.153)	0.716 (0.088)	0.810 (0.091)
	LSTM	0.786 (0.080)	0.804 (0.071)	0.840 (0.059)	0.860 (0.047)	0.870 (0.054)	0.884 (0.061)
	biLSTM	**0.911** (0.056)	**0.844** (0.067)	**0.860** (0.057)	**0.882** (0.047)	**0.875** (0.052)	**0.906** (0.054)
Precision	RF	0.680 (0.184)	0.670 (0.113)	0.654 (0.069)	0.636 (0.088)	0.675 (0.036)	0.719 (0.046)
	MLP	0.650 (0.214)	0.630 (0.079)	0.659 (0.069)	0.627 (0.073)	0.659 (0.040)	0.715 (0.045)
	SVM	0.733 (0.163)	0.649 (0.090)	0.707 (0.083)	0.684 (0.093)	0.692 (0.042)	0.725 (0.046)
	LSTM	0.682 (0.135)	0.692 (0.130)	0.706 (0.103)	0.683 (0.120)	0.710 (0.106)	0.727 (0.071)
	biLSTM	**0.788** (0.149)	**0.721** (0.132)	**0.722** (0.103)	**0.705** (0.118)	**0.752** (0.107)	**0.749** (0.067)
BACC	RF	0.781 (0.073)	0.776 (0.068)	0.761 (0.078)	0.768 (0.083)	0.770 (0.042)	0.831 (0.042)
	MLP	0.750 (0.085)	0.768 (0.057)	0.774 (0.069)	0.759 (0.067)	0.761 (0.049)	0.822 (0.046)
	SVM	0.787 (0.093)	0.775 (0.065)	0.787 (0.080)	0.784 (0.080)	0.787 (0.046)	0.831 (0.047)
	LSTM	0.822 (0.051)	0.828 (0.051)	0.847 (0.044)	0.840 (0.054)	0.848 (0.042)	0.860 (0.046)
	biLSTM	**0.906** (0.041)	**0.854** (0.048)	**0.861** (0.042)	**0.858** (0.051)	**0.867** (0.045)	**0.878** (0.040)
HSS	RF	0.548 (0.152)	0.541 (0.124)	0.516 (0.130)	0.516 (0.150)	0.536 (0.072)	0.639 (0.076)
	MLP	0.492 (0.183)	0.517 (0.096)	0.537 (0.118)	0.500 (0.119)	0.516 (0.083)	0.624 (0.081)
	SVM	0.585 (0.178)	0.534 (0.110)	0.575 (0.141)	0.561 (0.148)	0.568 (0.080)	0.640 (0.082)
	LSTM	0.612 (0.128)	0.624 (0.125)	0.656 (0.105)	0.633 (0.131)	0.656 (0.108)	0.682 (0.093)
	biLSTM	**0.769** (0.124)	**0.670** (0.123)	**0.682** (0.103)	**0.667** (0.126)	**0.702** (0.112)	**0.717** (0.083)
TSS	RF	0.562 (0.146)	0.551 (0.136)	0.523 (0.155)	0.536 (0.165)	0.541 (0.084)	0.662 (0.084)
	MLP	0.501 (0.171)	0.536 (0.114)	0.549 (0.139)	0.519 (0.134)	0.522 (0.097)	0.644 (0.093)
	SVM	0.574 (0.186)	0.551 (0.131)	0.573 (0.161)	0.568 (0.161)	0.574 (0.091)	0.661 (0.093)
	LSTM	0.645 (0.103)	0.657 (0.102)	0.695 (0.088)	0.680 (0.109)	0.697 (0.085)	0.720 (0.091)
	biLSTM	**0.812** (0.081)	**0.708** (0.096)	**0.722** (0.073)	**0.715** (0.103)	**0.733** (0.091)	**0.756** (0.079)
WAUC	RF	0.619 (0.022)	0.601 (0.046)	0.577 (0.022)	0.579 (0.050)	0.599 (0.065)	0.729 (0.072)
	MLP	0.551 (0.015)	0.581 (0.013)	0.605 (0.042)	0.573 (0.051)	0.565 (0.024)	0.709 (0.035)
	SVM	0.633 (0.015)	0.597 (0.079)	0.630 (0.048)	0.618 (0.057)	0.625 (0.056)	0.730 (0.046)
	LSTM	0.708 (0.039)	0.725 (0.020)	0.757 (0.084)	0.744 (0.081)	0.761 (0.029)	0.782 (0.070)
	biLSTM	**0.895** (0.013)	**0.775** (0.041)	**0.799** (0.051)	**0.780** (0.069)	**0.796** (0.058)	**0.821** (0.063)

Goodfellow, I. J., Bengio, Y., & Courville, A. C. 2016, Deep Learning, Adaptive computation and machine learning (Cambridge, MA: MIT Press). http://www.deeplearningbook.org/

Harrison, R. A. 1995, A&A, 304, 585. https://ui.adsabs.harvard.edu/abs/1995A&A...304..585H

Heidke, P. 1926, Geografiska Annaler, 8, 301, doi: 10.1080/20014422.1926.11881138

Hochreiter, S., & Schmidhuber, J. 1997, Neural Computation, 9, 1735, doi: 10.1162/neco.1997.9.8.1735

Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483, doi: 10.1007/s11207-014-0516-8

Huang, N., Xu, Y., & Wang, H. 2018, Research Notes of the AAS, 2, 7, doi: 10.3847/2515-5172/aaa602

Huang, X., Wang, H., Xu, L., et al. 2018, ApJ, 856, 7, doi: 10.3847/1538-4357/aaae00

Inceoglu, F., Jeppesen, J. H., Kongstad, P., et al. 2018, ApJ, 861, 128, doi: 10.3847/1538-4357/aac81e

Jonas, E., Bobra, M., Shankar, V., Todd Hoeksema, J., & Recht, B. 2018, SoPh, 293, 48, doi: 10.1007/s11207-018-1258-9

Jordanova, V. K., Delzanno, G. L., Henderson, M. G., et al. 2018, Journal of Atmospheric and Solar-Terrestrial Physics, 177, 148, doi: 10.1016/j.jastp.2017.11.006

Kang, H., Yang, S.-H., Huang, J., & Oh, J. 2020, International Journal of Control Automation and Systems, 18, 3023, doi: 10.1007/s12555-019-0984-6

Table 7. Performance Comparison of RF, MLP, SVM, LSTM and biLSTM Based on the Random Division Scheme for the F_S Problem

		12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
Recall	RF	0.734 (0.105)	0.797 (0.117)	0.749 (0.147)	0.710 (0.116)	0.717 (0.089)	0.756 (0.104)
	MLP	0.683 (0.087)	0.769 (0.123)	0.752 (0.109)	0.689 (0.121)	0.702 (0.103)	0.728 (0.083)
	SVM	0.767 (0.107)	0.791 (0.101)	0.786 (0.104)	0.730 (0.120)	0.740 (0.103)	0.743 (0.086)
	LSTM	0.774 (0.098)	0.770 (0.125)	0.817 (0.108)	0.782 (0.120)	0.768 (0.106)	0.772 (0.103)
	biLSTM	**0.834** (0.096)	**0.837** (0.120)	**0.856** (0.107)	**0.837** (0.122)	**0.818** (0.101)	**0.815** (0.102)
Precision	RF	0.243 (0.117)	0.244 (0.103)	0.261 (0.096)	0.308 (0.089)	0.396 (0.190)	0.434 (0.162)
	MLP	0.225 (0.107)	0.243 (0.100)	0.269 (0.082)	0.291 (0.069)	0.363 (0.156)	0.366 (0.097)
	SVM	0.235 (0.104)	0.231 (0.086)	0.242 (0.067)	0.326 (0.099)	0.383 (0.166)	0.393 (0.113)
	LSTM	0.250 (0.114)	0.252 (0.097)	0.291 (0.092)	0.349 (0.108)	0.413 (0.189)	0.443 (0.165)
	biLSTM	**0.279** (0.131)	**0.275** (0.107)	**0.306** (0.099)	**0.377** (0.119)	**0.483** (0.174)	**0.476** (0.173)
BACC	RF	0.770 (0.073)	0.774 (0.058)	0.758 (0.087)	0.760 (0.054)	0.760 (0.083)	0.795 (0.049)
	MLP	0.742 (0.065)	0.764 (0.058)	0.766 (0.054)	0.746 (0.056)	0.748 (0.069)	0.770 (0.045)
	SVM	0.783 (0.069)	0.770 (0.057)	0.764 (0.071)	0.772 (0.056)	0.770 (0.069)	0.783 (0.046)
	LSTM	0.791 (0.068)	0.772 (0.062)	0.799 (0.053)	0.801 (0.056)	0.787 (0.069)	0.804 (0.049)
	biLSTM	**0.825** (0.068)	**0.809** (0.059)	**0.821** (0.053)	**0.832** (0.057)	**0.841** (0.059)	**0.830** (0.049)
HSS	RF	0.284 (0.151)	0.274 (0.129)	0.284 (0.128)	0.327 (0.102)	0.390 (0.196)	0.442 (0.153)
	MLP	0.255 (0.138)	0.269 (0.125)	0.295 (0.098)	0.306 (0.089)	0.358 (0.166)	0.379 (0.110)
	SVM	0.280 (0.138)	0.260 (0.111)	0.267 (0.101)	0.350 (0.110)	0.387 (0.174)	0.409 (0.122)
	LSTM	0.298 (0.149)	0.284 (0.122)	0.330 (0.110)	0.385 (0.115)	0.418 (0.189)	0.455 (0.155)
	biLSTM	**0.340** (0.165)	**0.321** (0.131)	**0.354** (0.116)	**0.426** (0.124)	**0.515** (0.158)	**0.500** (0.162)
TSS	RF	0.540 (0.145)	0.548 (0.116)	0.516 (0.174)	0.519 (0.109)	0.521 (0.166)	0.589 (0.098)
	MLP	0.484 (0.130)	0.527 (0.116)	0.533 (0.108)	0.493 (0.111)	0.497 (0.138)	0.540 (0.090)
	SVM	0.566 (0.138)	0.540 (0.113)	0.528 (0.142)	0.545 (0.111)	0.539 (0.138)	0.567 (0.093)
	LSTM	0.581 (0.137)	0.544 (0.125)	0.598 (0.106)	0.601 (0.112)	0.574 (0.138)	0.607 (0.097)
	biLSTM	**0.651** (0.136)	**0.617** (0.119)	**0.641** (0.106)	**0.664** (0.114)	**0.682** (0.118)	**0.660** (0.097)
WAUC	RF	0.591 (0.057)	0.600 (0.021)	0.561 (0.016)	0.565 (0.038)	0.577 (0.043)	0.655 (0.057)
	MLP	0.528 (0.068)	0.577 (0.043)	0.578 (0.069)	0.543 (0.065)	0.547 (0.054)	0.597 (0.085)
	SVM	0.624 (0.025)	0.586 (0.027)	0.580 (0.032)	0.602 (0.039)	0.590 (0.069)	0.617 (0.065)
	LSTM	0.634 (0.052)	0.595 (0.048)	0.665 (0.026)	0.666 (0.027)	0.625 (0.028)	0.673 (0.062)
	biLSTM	**0.712** (0.068)	**0.680** (0.082)	**0.710** (0.057)	**0.730** (0.053)	**0.753** (0.031)	**0.732** (0.060)

Kawabata, Y., Iida, Y., Doi, T., et al. 2018, ApJ, 869, 99, doi: 10.3847/1538-4357/aaebfc

Kilpua, E., Koskinen, H. E. J., & Pulkkinen, T. I. 2017, Living Reviews in Solar Physics, 14, 5, doi: 10.1007/s41116-017-0009-6

Kruskal, J. B. 1964, Psychometrika, 29, 115, doi: 10.1007/BF02289694

LeCun, Y., Bengio, Y., & Hinton, G. 2015, Nature, 521, 436, doi: 10.1038/nature14539

Leka, K. D., & Barnes, G. 2003, ApJ, 595, 1277, doi: 10.1086/377511

Lin, J., & Forbes, T. G. 2000, J. Geophys. Res., 105, 2375, doi: 10.1029/1999JA900477

Liu, C., Deng, N., Wang, J. T. L., & Wang, H. 2017, ApJ, 843, 104, doi: 10.3847/1538-4357/aa789b

Liu, H. 2020, Ph.D. Dissertation, New Jersey Institute of Technology. https://digitalcommons.njit.edu/dissertations/1481

Liu, H., Liu, C., Wang, J. T. L., & Wang, H. 2019, ApJ, 877, 121, doi: 10.3847/1538-4357/ab1b3c

Liu, H., Liu, C., Wang, J. T. L., & Wang, H. 2020, ApJ, 890, 12, doi: 10.3847/1538-4357/ab6850

Marzban, C. 2004, Weather and Forecasting, 19, 1106, doi: 10.1175/825.1

Table 8. Probabilistic Forecasting Results of RF, MLP, SVM, LSTM and biLSTM With and Without Calibration Based on the Random Division Scheme for the FC_S and F_S Problems Respectively

			12 hr	24 hr	36 hr	48 hr	60 hr	72 hr
FC_S	BS	RF	0.268 (0.068)	0.279 (0.076)	0.305 (0.102)	0.300 (0.119)	0.232 (0.043)	0.305 (0.040)
		RF+C	0.226 (0.057)	0.246 (0.066)	0.260 (0.087)	0.256 (0.102)	0.272 (0.038)	0.269 (0.036)
		MLP	0.311 (0.101)	0.303 (0.084)	0.339 (0.118)	0.312 (0.119)	0.303 (0.046)	0.311 (0.066)
		MLP+C	0.265 (0.087)	0.259 (0.071)	0.288 (0.101)	0.266 (0.102)	0.283 (0.040)	0.293 (0.058)
		SVM	0.261 (0.073)	0.264 (0.089)	0.271 (0.121)	0.261 (0.127)	0.285 (0.065)	0.282 (0.043)
		SVM+C	0.219 (0.061)	0.241 (0.076)	0.252 (0.103)	0.253 (0.109)	0.249 (0.058)	0.231 (0.040)
		LSTM	0.258 (0.041)	0.263 (0.041)	0.278 (0.035)	0.272 (0.044)	0.257 (0.034)	0.268 (0.036)
		LSTM+C	0.220 (0.037)	0.224 (0.037)	0.236 (0.031)	0.232 (0.040)	0.238 (0.031)	0.235 (0.033)
		biLSTM	0.214 (0.021)	0.230 (0.034)	0.203 (0.037)	0.214 (0.040)	0.193 (0.024)	0.185 (0.012)
		biLSTM+C	**0.183** (0.019)	**0.195** (0.027)	**0.153** (0.028)	**0.182** (0.035)	**0.164** (0.020)	**0.157** (0.011)
	BSS	RF	0.464 (0.138)	0.429 (0.154)	0.402 (0.207)	0.408 (0.235)	0.354 (0.099)	0.368 (0.092)
		RF+C	0.549 (0.118)	0.516 (0.132)	0.490 (0.176)	0.495 (0.201)	0.453 (0.088)	0.460 (0.084)
		MLP	0.381 (0.205)	0.399 (0.169)	0.320 (0.243)	0.378 (0.241)	0.338 (0.096)	0.221 (0.139)
		MLP+C	0.474 (0.175)	0.488 (0.143)	0.422 (0.209)	0.469 (0.207)	0.439 (0.084)	0.335 (0.122)
		SVM	0.476 (0.154)	0.431 (0.176)	0.433 (0.248)	0.386 (0.258)	0.415 (0.133)	0.409 (0.096)
		SVM+C	0.560 (0.129)	0.517 (0.151)	0.509 (0.210)	0.488 (0.222)	0.505 (0.120)	0.517 (0.087)
		LSTM	0.492 (0.085)	0.478 (0.082)	0.445 (0.074)	0.455 (0.092)	0.452 (0.077)	0.430 (0.079)
		LSTM+C	0.566 (0.075)	0.556 (0.072)	0.529 (0.067)	0.535 (0.086)	0.533 (0.067)	0.516 (0.070)
		biLSTM	0.573 (0.053)	0.545 (0.070)	0.596 (0.078)	0.569 (0.087)	0.610 (0.056)	0.627 (0.036)
		biLSTM+C	**0.635** (0.047)	**0.614** (0.060)	**0.696** (0.058)	**0.633** (0.077)	**0.668** (0.047)	**0.683** (0.032)
F_S	BS	RF	0.284 (0.076)	0.288 (0.061)	0.372 (0.067)	0.320 (0.130)	0.307 (0.094)	0.310 (0.052)
		RF+C	0.272 (0.099)	0.276 (0.087)	0.312 (0.056)	0.268 (0.109)	0.274 (0.121)	0.281 (0.101)
		MLP	0.332 (0.114)	0.329 (0.072)	0.394 (0.080)	0.352 (0.080)	0.316 (0.145)	0.332 (0.147)
		MLP+C	0.283 (0.098)	0.278 (0.063)	0.336 (0.071)	0.279 (0.123)	0.299 (0.124)	0.305 (0.124)
		SVM	0.276 (0.067)	0.279 (0.063)	0.298 (0.059)	0.276 (0.054)	0.295 (0.058)	0.290 (0.045)
		SVM+C	0.236 (0.059)	0.249 (0.055)	0.264 (0.050)	0.267 (0.045)	0.291 (0.049)	0.248 (0.038)
		LSTM	0.277 (0.065)	0.279 (0.059)	0.285 (0.051)	0.282 (0.080)	0.273 (0.066)	0.289 (0.046)
		LSTM+C	0.236 (0.056)	0.231 (0.051)	0.242 (0.044)	0.246 (0.068)	0.247 (0.058)	0.247 (0.040)
		biLSTM	0.260 (0.054)	0.247 (0.048)	0.260 (0.043)	0.265 (0.046)	0.253 (0.085)	0.237 (0.089)
		biLSTM+C	**0.221** (0.046)	**0.209** (0.041)	**0.221** (0.037)	**0.223** (0.039)	**0.214** (0.073)	**0.201** (0.076)
	BSS	RF	0.390 (0.133)	0.365 (0.077)	0.397 (0.071)	0.342 (0.139)	0.330 (0.164)	0.324 (0.162)
		RF+C	0.463 (0.158)	0.470 (0.140)	0.482 (0.089)	0.424 (0.172)	0.440 (0.194)	0.419 (0.167)
		MLP	0.289 (0.205)	0.348 (0.146)	0.215 (0.170)	0.260 (0.191)	0.264 (0.208)	0.200 (0.153)
		MLP+C	0.388 (0.187)	0.450 (0.127)	0.331 (0.150)	0.366 (0.186)	0.357 (0.204)	0.304 (0.165)
		SVM	0.447 (0.145)	0.420 (0.134)	0.403 (0.129)	0.378 (0.109)	0.404 (0.117)	0.391 (0.097)
		SVM+C	0.529 (0.126)	0.493 (0.117)	0.492 (0.109)	0.461 (0.091)	0.420 (0.101)	0.507 (0.083)
		LSTM	0.450 (0.139)	0.467 (0.128)	0.436 (0.103)	0.422 (0.145)	0.442 (0.139)	0.421 (0.101)
		LSTM+C	0.531 (0.118)	0.536 (0.110)	0.511 (0.088)	0.500 (0.146)	0.461 (0.121)	0.506 (0.086)
		biLSTM	0.484 (0.112)	0.505 (0.105)	0.489 (0.089)	0.466 (0.097)	0.428 (0.154)	0.432 (0.166)
		biLSTM+C	**0.592** (0.095)	**0.581** (0.090)	**0.566** (0.075)	**0.550** (0.082)	**0.504** (0.165)	**0.613** (0.074)

Meinshausen, N., & Bühlmann, P. 2010, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72, 417, doi: 10.1111/j.1467-9868.2010.00740.x

Moore, R. L., Falconer, D. A., & Sterling, A. C. 2012, ApJ, 750, 24, doi: 10.1088/0004-637X/750/1/24

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. 2018, ApJ, 858, 113, doi: 10.3847/1538-4357/aab9a7

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825. https://jmlr.org/papers/v12/pedregosa11a.html

Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, The Solar Dynamics Observatory (SDO) (New York, NY: Springer US), 3–15, doi: 10.1007/978-1-4614-3673-7_2

Reames, D. V., Ng, C. K., & Tylka, A. J. 2013, SoPh, 285, 233, doi: 10.1007/s11207-012-0038-1

Roeder, J. L., & Jordanova, V. K. 2020, in Ring Current Investigations, ed. V. K. Jordanova, R. Ilie, & M. W. Chen (Elsevier), 245–269, doi: 10.1016/B978-0-12-815571-4.00008-1

Rosenblatt, F. 1958, The Perceptron: A Theory of Statistical Separability in Cognitive Systems (Cornell Aeronautical Laboratory, Inc., Rep. No. VG-1196-G-1. U.S), 268, doi: 10.1037/h0042519

Sager, T. W., & Thisted, R. A. 1982, The Annals of Statistics, 10, 690 , doi: 10.1214/aos/1176345865

Schou, J., Borrero, J. M., Norton, A. A., et al. 2012, SoPh, 275, 327, doi: 10.1007/s11207-010-9639-8

Schrijver, C. J. 2007, ApJL, 655, L117, doi: 10.1086/511857

Schuster, M., & Paliwal, K. 1997, IEEE Transactions on Signal Processing, 45, 2673, doi: 10.1109/78.650093

Siami-Namini, S., Tavakoli, N., & Namin, A. S. 2019, in 2019 IEEE International Conference on Big Data (Big Data), 3285–3292, doi: 10.1109/BigData47090.2019.9005997

SunPy Community, T., Mumford, S. J., Christe, S., et al. 2015, Computational Science and Discovery, 8, 014009, doi: 10.1088/1749-4699/8/1/014009

Tibshirani, R. 1996, Journal of the Royal Statistical Society: Series B (Methodological), 58, 267, doi: 10.1111/j.2517-6161.1996.tb02080.x

van Driel-Gesztelyi, L., & Green, L. M. 2015, Living Reviews in Solar Physics, 12, 1, doi: 10.1007/lrsp-2015-1

Wang, X., Chen, Y., Toth, G., et al. 2020, ApJ, 895, 3, doi: 10.3847/1538-4357/ab89ac

Webb, D. F., & Howard, T. A. 2012, Living Reviews in Solar Physics, 9, 3, doi: 10.12942/lrsp-2012-3

Wilks, D. S. 2010, Quarterly Journal of the Royal Meteorological Society, 136, 2109, doi: 10.1002/qj.709

Yashiro, S., & Gopalswamy, N. 2008, Proceedings of the International Astronomical Union, 4, 233–243, doi: 10.1017/S1743921309029342