Space Weather

RESEARCH ARTICLE
10.1029/2020SW002440

Solar Flare Intensity Prediction With Machine Learning Models

Zhenbang Jiao[1], Hu Sun[1], Xiantong Wang[2], Ward Manchester[2], Tamas Gombosi[2], Alfred Hero[1,3], and Yang Chen[1,4]

[1]Department of Statistics, University of Michigan, Ann Arbor, MI, USA, [2]Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI, USA, [3]Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, [4]Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI, USA

Key Points:
- We develop deep learning models to predict solar flare intensity values instead of flare classes from SHARP parameters in SDO/HMI data set directly
- We use time-series information from both flaring time and nonflaring time in our model
- As opposed to solar flare classification, directly predicting solar flare intensity gives more detailed information about every occurrence of flares of each class

Correspondence to:
Y. Chen,
ychenang@umich.edu

Citation:
Jiao, Z., Sun, H., Wang, X., Manchester, W., Gombosi, T., Hero, A., & Chen, Y. (2020). Solar flare intensity prediction with machine learning models. *Space Weather*, *18*, e2020SW002440. https://doi.org/10.1029/2020SW002440

Received 2 JAN 2020
Accepted 11 MAY 2020
Accepted article online 15 MAY 2020

Abstract We develop a mixed long short-term memory (LSTM) regression model to predict the maximum solar flare intensity within a 24-hr time window 0–24, 6–30, 12–36, and 24–48 hr ahead of time using 6, 12, 24, and 48 hr of data (predictors) for each Helioseismic and Magnetic Imager (HMI) Active Region Patch (HARP). The model makes use of (1) the Space-Weather HMI Active Region Patch (SHARP) parameters as predictors and (2) the exact flare intensities instead of class labels recorded in the Geostationary Operational Environmental Satellites (GOES) data set, which serves as the source of the response variables. Compared to solar flare classification, the model offers us more detailed information about the exact maximum flux level, that is, intensity, for each occurrence of a flare. We also consider classification models built on top of the regression model and obtain better results in solar flare classifications as compared to Chen et al. (2019, https://doi.org/10.1029/2019SW002214). Our results suggest that the most efficient time period for predicting the solar activity is within 24 hr before the prediction time using the SHARP parameters and the LSTM model.

1. Introduction

Space weather involves the dynamical processes of the Sun-Earth system that may affect human life and technology. The most destructive consequences of space weather, ranging from electric power disruptions to radiation hazards for astronauts, are due to energetic solar eruptions, producing both magnetic disturbances in the solar wind known as coronal mass ejections (CMEs) and intense electromagnetic radiation known as solar flares.

Given their destructive capability, the predictions of energetic space weather events are critical for safeguarding our technological infrastructure. Extreme space storms—those that could significantly degrade critical infrastructure—could disable large portions of the electrical power grid, resulting in cascading failures that would affect key services such as water supply, health care, and transportation. The threat assessment report by the Lloyd's insurance company (Maynard et al., 2013) concludes that extreme events could cause $2.6 trillion in damage with a recovery time of months. An earlier report by the National Research Council (Baker et al., 2009) arrived at similar conclusions.

While there are known precursors to these eruptions, accurate predictions of their occurrence remain very difficult. The current space weather forecasting based on physical models is far from reliable: The forecasting window is only minutes away from the current time point, and the accuracy is low. Previous work has established that solar eruptions are all associated with highly nonpotential magnetic fields that store the necessary free energy. The most energetic flares come from very localized intense kilogauss photospheric fields known as active regions (Forbes, 2000; Schrijver, 2009). Measurement of these fields was greatly increased by the advent of the Helioseismic and Magnetic Imager (HMI) instrument on the Solar Dynamics Observatory (SDO) launched on February 2010. HMI provides vast quantities of data in the form of high-cadence high-resolution vector magnetograms. These data are subdivided into HMI Active Region Patches (HARPs), which correspond to localized regions of intense magnetic fields. While HARPs are very similar to National Oceanic and Atmospheric Administration (NOAA) active regions, they frequently define different spatial regions. Parameters relevant to solar eruptions are calculated from the HARP vector magnetic fields and saved with the data files which are designated as Space-Weather HMI Active Region Patches or SHARPs (Bobra et al., 2014).

Currently, over 7,000 HARPs have been recorded, each one with full vector data saved on a 12-min cadence for a period of approximately 14 days required to rotate across the disk. How to make the best use of the large amount of data available to provide reliable real-time forecasting of space weather events is one of the major questions for scientists in the field. Recently, data-driven approaches are gaining attention in the space science community with much more data becoming available. Scientists have adopted different machine learning algorithms to perform various space weather prediction tasks, including the solar flare classification using the SDO/HMI SHARP parameters and other data sets; see Barnes et al. (2016), Leka and Barnes (2018), Liu et al. (2019), Camporeale (2019), Leka et al. (2019a), and Leka et al. (2019b) for reviews and references therein. Among all the papers mentioned, Liu et al. (2019) also used the Geostationary Operational Environmental Satellites (GOES) data set and adopted the long short-term memory (LSTM) technique to predict solar flares. In contrast, in this paper, we propose a different mixed LSTM model, and we consider not only classification but also regression to predict the exact intensities rather than the labels of the solar flares. Moreover, our data preprocessing gives a new way of defining response variables and takes quiet time data into consideration.

Chen et al. (2019) showed that the time series of SHARP parameters from the SDO/HMI data provide useful information for distinguishing strong solar flares of M/X class from weak flares of A/B class roughly 24 hr prior to the flare event. These SHARP parameters are derived from the HMI images based on physically meaningful quantities of the active regions where the flares emerge from; see Bobra et al. (2014) for detailed descriptions of these features. To make the task of binary classification manageable, Chen et al. (2019) only considered the B and M/X flares, ignoring the more prevalent C flares. This design is due to the consideration that flare classes are arbitrarily categorized based on a continuous logarithmic scale of flare intensity (radiant power level), thus strong C flares are essentially indistinguishable from weak M flares.

Figure 1 shows the flare history (B/C/M/X classes) for two HARPs (377 and 746) and time evolution of two important SHARP parameters, TOTUSJH and SAVNCPP, for a period of 10 days (labeled on the x axis). Specifically, TOTUSJH stands for total unsigned current helicity, and SAVNCPP stands for sum of the modulus of the net current per polarity. We can see that many incidences of C flares accompany a strong flare (of M/X class) and that the SHARP parameters evolve in continuous but locally stochastic ways during the energy buildup and release stages of strong flares. Therefore, it is important to consider the entire time series with flares of all classes, especially the highly prevalent C flares, when training machine learning models for flare prediction as opposed to only the time point where a weak (B) or strong (M/X) flare occurs as is done in Chen et al. (2019).

As found in the GOES data set, flare events occur sparsely, at irregular intervals, and at highly varying intensity levels, including long gaps between events, all of which present a unique challenge in the data analysis. We note that due to the fact that the amount of information contained in the observed data is limited, the inferential objective should be geared toward extracting the maximum amount of *available* information and avoiding overinterpreting the data. Instead of seeking to model the flare intensity in continuous time for every time point, we model aggregated quantities instead, for example, the maximum flare intensity within a fixed length time window (such as $\pm$12 hr). In this way, we attach an intensity value to every data point that has a recorded flare in the neighboring $\pm$12-hr time window. For the other time points, we define them as being "quiet" locally with an indicator function attached to it. We will explain the details of this data preparation process in section 2.1. In our proposed prediction model, we are able to predict the maximum flare intensity level within a fixed length time window T hours in the future, where T can be specified to a desired value such as 12 or 24 hr, using the time series of SHARP parameters in the past. As a by-product, we can classify the predicted events into strong or weak flares according to the flare level definitions.

2. Methodology

We provide a detailed description of the data preprocessing pipeline in section 2.1. A mixed LSTM regression model (Hochreiter & Schmidhuber, 1997) that can directly predict the solar flare intensity is introduced in section 2.2, including the model structure and a novel loss function. Section 2.3 covers three binary classification models based on the mixed LSTM regression model. They all try to distinguish the M and X flares from other flares (including or excluding the C flares) by making use of the predicted intensities given by the regression model.

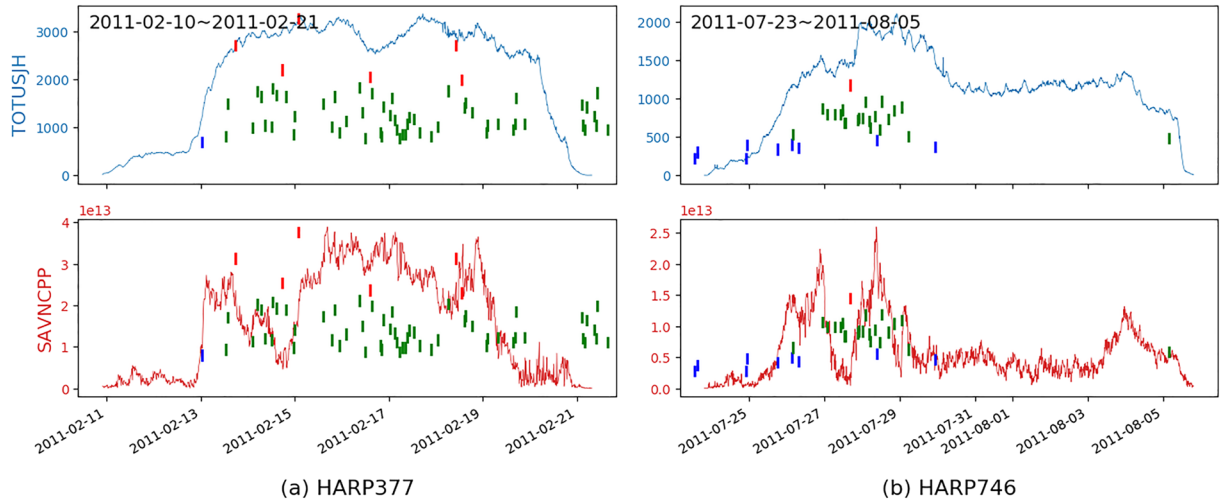

Figure 1. Examples of physical parameters derived from two HARPs, (a) 377 and (b) 746. The blue and red curves show the time variation of TOTUSJH and SAVNCPP quantities, respectively. Here, TOTUSJH stands for total unsigned current helicity, and SAVNCPP stands for sum of the modulus of the net current per polarity. Each small vertical line represents a recorded flare event. The height of the line is proportional to the log scale flare intensity, while red, green, and blue represent M/X flare, C flare, and B flare, respectively.

2.1. Data Preparation

The machine learning models that we aim to train are prediction models, which require two sources of input data: the feature set (a.k.a. predictors) and the response variables. In this section, we give the details of the data sources and how we prepare the data for training and testing the machine learning models.

For response variables, we use flare events recorded in the GOES data set ranging from 1 May 2010 to 20 June 2018. Within this time range, there are a total of 12,012 recorded flares. See flare-event-only data set in Figure 2 for the distribution of the flare events in GOES data set. Note that the theoretical distribution of the flare events should be a power law distribution. The reduced number of recorded flares in lower energy levels is because events are lost in the background and go undetected. Therefore, the observed distribution is different from the theoretical distribution, and we are focused on the observed information in this paper.

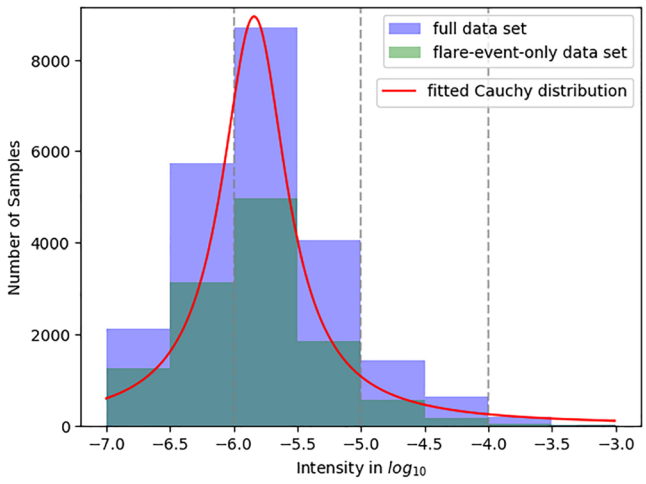

Figure 2. The distribution of nonquiet samples' flare intensities (I) in flare-event-only data set and full data set, where flare-event-only data set only takes flare intensities recorded on GOES data set as response variables. The definition of full data set can be seen in section 2.1.1. Red line is the fitted Cauchy distribution with location parameter x_0 = -5.84 and scale parameter $\gamma = 0.31$.

For the source of data for features/predictors, we consider data from 860 HARPs. For the chosen time period (1 May 2010 to 20 June 2018), there are approximately 7,000 HARPs, many occurring without flares. From these, in order to maintain the quality of the data, we downselect the HARPs to a group of 860 based on the criteria that (1) the longitude of the HARP should be within the range of ±68° from Sun central meridian, to avoid projection effects (see Bobra & Couvidat, 2015 and Chen et al., 2019) and (2) the missing SHARP parameters should be fewer than 5% of all in the HARP, to make sure that the missing data is not significantly large to cause any bias in model training.

For each HARP, there is a time series of vector magnetograms with 12-min cadence. Here we consider the time series as a video with one frame every 12 min. We use the SHARP parameters, which are scalar variables derived from the full photospheric vector magnetic field. The SHARP parameters are calculated over the magnetogram of the each frame; see Bobra et al. (2014) for a detailed description of the calculations. Of all the SHARP parameters, we use USFLUX, MEANGAM, MEANGBT, MEANGBZ, MEANGBH, MEANJZD, TOTUSJZ, MEANALP, MEANJZH, TOTUSJH, ABSNJZH, SAVNCPP, MEANPOT, TOTPOT, MEANSHR, SHRGT45, SIZE, SIZE_ACR, NACR, and NPIX in our study (see the definitions of

Table 1
List of SHARP Parameters and Brief Descriptions

Parameter	Description
TOTUSJH:	Total unsigned current helicity
TOTUSJZ:	Total unsigned vertical current
SAVNCPP:	Sum of the modulus of the net current per polarity
USFLUX:	Total unsigned flux
ABSNJZH:	Absolute value of the net current helicity
TOTPOT:	Proxy for total photospheric magnetic free energy density
SIZE ACR:	Deprojected area of active pixels (B_z magnitude larger than noise threshold) on image in microhemisphere (defined as one millionth of half the surface of the Sun)
NACR:	The number of strong LoS magnetic-field pixels in the patch
MEANPOT:	Proxy for mean photospheric excess magnetic energy density
SIZE:	Projected area of the image in microhemispheres
MEANJZH:	Current helicity (Bz contribution)
SHRGT45:	Fraction of area with shear $> 45°$
MEANSHR:	Mean shear angle
MEANJZD:	Vertical current density
MEANALP:	Characteristic twist parameter, α
MEANGBT:	Horizontal gradient of total field
MEANGAM:	Mean angle of field from radial
MEANGBZ:	Horizontal gradient of vertical field
MEANGBH:	Horizontal gradient of horizontal field
NPIX:	Number of pixels within the patch

these parameters in Table 1). Therefore, each frame corresponds to one vector magnetogram and a 20×1 SHARP vector. Each HARP corresponds to a data matrix with 20 columns and "number of frames (vector magnetograms)" rows. These data are provided by the Stanford Joint Science Operations Center (see https://jsoc.stanford.edu).

2.1.1. Response Variable

Since some of the flares recorded in the GOES data set happened in HARPs that are not recorded in the filtered JSOC data, we consider 10,349 out of the total 12,012 flares recorded in the GOES data set during the time range indicated on Table 2. Moreover, the flares recorded in the GOES data set are listed by NOAA active region numbers, while the corresponding photospheric magnetic field is identified with HARP patches, which use different criteria to identify and group the strong field regions. Consequently, there is the potential issue of a single HARP corresponding to multiple active regions; in fact, roughly 20% of SHARP patches include components from multiple active regions. This problem has been acknowledged in Chen et al. (2019), and more details can be found therein. In this paper, we do not address this potential problem caused by the data but focus on the methods for modeling. We speculate that this potential problem of mismatch of SHARP and GOES data may or may not result in biases for prediction models while might incur loss of statistical efficiency due to the extra noise brought in.

In order to make maximum use of the data, we consider not only the class of each flare but also the exact value of the flare intensity which is defined as the peak flux in watts per square meter (W/m^2) of soft X-rays with wavelengths 100 to 800 pm. Moreover, since the flare intensity spans orders of magnitude, we take the $\log_{10}$ transform (see Table 3) in order to better handle the extreme values, X and M flares. All flare intensities mentioned later are $\log_{10}$ scale intensities if not further specified.

After performing the data processing as described above, there are over 10,000 flares identified from a time history of X-ray intensity levels. However, considering only the peak intensity level recorded at a given time point as in Chen et al. (2019), there are some limitations, stated below.

1. Most of the M and X flare events are accompanied by much more frequent C flares. If we simply assign the response variable based on flares' peak times, two flares happening adjacent to each other with totally different intensities can have a large amount of overlapping training data (time series). Two observations with similar training data but quite different response variables would confuse the model.
2. Even though there are over 10,000 flare records in GOES data set, they are not all in the recorded range of the 860 HARP videos. Also, the number of the strong flares which we care the most are limited (see Table 2). Besides, some of the HARP videos are not suitable for use in training machine learning models due to large amounts of missing entries in the SHARP parameters. Therefore, the effective number of flare events that we can use for training/testing the machine learning model is not as large as expected.
3. The recorded flares only occupy a very small fraction of the time series of observations, that is, the SHARP parameters. Those time points without a recorded flare might be an unrecorded weak flare near a stronger one or most likely a "flare-free" time point. Considering these time points as contrasts to the

Table 2
The Number of X/M/C/B Flares Recorded in Each Year in the GOES Data Set During the Time Range 1 May 2010 to 20 June 2018

Class/Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
X	0	8	5	12	15	2	0	4	0	46
M	8	84	110	90	169	128	7	37	0	633
C	64	788	906	1,105	1,231	1,194	244	225	11	5,768
B	512	519	398	418	94	428	722	606	205	3,902

Table 3
Transformation From Flares Class to Continuous Intensity Values We Adopt

Flare class	Peak flux range (W/m^2)	$\log_{10}$ intensity
X	$\geq 10^{-4}$	≥ -4
M	10^{-5} to 10^{-4}	-5 to -4
C	10^{-6} to 10^{-5}	-6 to -5
B	10^{-7} to 10^{-6}	-7 to -6

$\log_{10}$

time points with flares can help the model better distinguish the strong flares from the others. Therefore, discarding this piece of information would impair the performance of the prediction model.

Therefore, in order to overcome these drawbacks, we propose the following definition of response variables in our prediction model: For each frame, we define its real-time intensity as the maximum flare intensity that happened within a 24-hr time window (12 hr before and 12 hr after). In other words, instead of focusing on each recorded flare in GOES data set, we only care about the largest flare that happened in each frame's 24-hr time window. By applying this new mechanism, we can assign each frame a response variable. Correspondingly, the new data set is called "full data set" (see the distribution of the flares in the constructed full data set as compared to the flare-event-only data set in Figure 2). As a result, the nonquiet sample size of the full data set is over two times larger as compared to the flare-event-only data set, 22,928 as opposed to 10,349. Plus, the response variables of those C flares happening next to strong flares (M or X) are redefined as high intensities, which is certainly more reasonable for model training. Most importantly, this mechanism more accurately portrays the processes of solar activities: Instead of being single-time-point incidences, they are processes of extended time evolution.

A natural question is how to deal with the frames where there is no flare recorded in the 24-hr time window. We define one more binary response variable to denote the "flaring" or "nonflaring" of the 24-hr time window—1 means there is at least one flare (M/X/C/B-class) recorded in the GOES data set within the 24-hr window, while 0 means no flare recorded in the GOES data set within the window.

To recap, for each frame, we assign it a two-dimensional response variable; the first dimension Q corresponds to the "local quietness" or "local nonquietness" (Boolean, 1 for having a flare event within the 24-hr window and 0 for not having a flare event within the 24-hr window), while the second dimension I stands for its real-time intensity on the $\log_{10}$ scale (continuous). Specifically, if a sample has $Q = 0$, then we annotate the second dimension of its response variable as N/A (see Table 4). An example of how we define the response variable $[Q, I]$ for HARP 377 is shown in Figure 3.

2.1.2. Input Data Preprocessing Pipeline

A detailed diagram of how we prepare the raw data for machine learning is shown in Figure 4. We briefly describe it here. Suppose we aim to train a model that uses m hours of SHARP parameters to predict the maximum flare intensity in the 24-hr window beginning at n hours after. Since the time cadence of our data is 12 min, there are five observed frames (magnetograms) at each hour. Each video needs to contain $5 \times (m + n + 24)$ consecutive frames to have at least one sample available. We take samples every 2 hr (10 frames), a reasonable step size which is neither too long to capture the detailed behaviors of the HARP nor so short that it causes oversampling of the time series. We take HARP 394 as an example. There are 1,334 frames in total. The training samples include frame 0 ~ frame $5m - 1$, frame 10 ~ frame $5m + 9$, ..., frame $10k$ ~ frame $5m + 10k - 1$, ... Correspondingly, the response variables include the maximum flare intensities recorded within frame $5(m + n)$ ~ frame $5(m + n + 24) - 1$, frame $5(m + n) + 10$ ~ frame $5(m + n + 24) + 9$, ..., frame $5(m + n) + 10k$ ~ frame $5(m + n + 24) + 10k - 1$, ..., where $k = 0,1,2...$ and $5(m + n + 24) + 10k - 1 < 1334$.

We split the training and testing data by years in order to avoid information leaking. Since all the recorded data range from 2010 to 2018, we have that roughly 63% of flares happened before 2015 (6,536 out of 10,349). We note that the corresponding sample size as obtained by the data preparation described above has a similar flare rate. Each HARP only has one video, so no HARP is divided in both the training and testing sets. In this study, we split all flares that happened before 1 January 2015 into the training set and the rest into the testing set. After splitting the data into training/testing samples, we normalize all the data by subtracting the mean and dividing by the standard deviation computed from the training data (Hastie et al., 2009, chapter 7.10). No information from the testing data is used in the normalization step.

Table 4
Examples of How We Define Response Variables Given the Flare Labels

Label	Response variable ($[Q,I]$)
M1.5	[1, −4.824]
X1.6	[1, −3.796]
C7.2	[1, −5.143]
Quiet	[0, N/A]

Note. Quiet stands for one quiet sample. See section 2.1.1 for details.

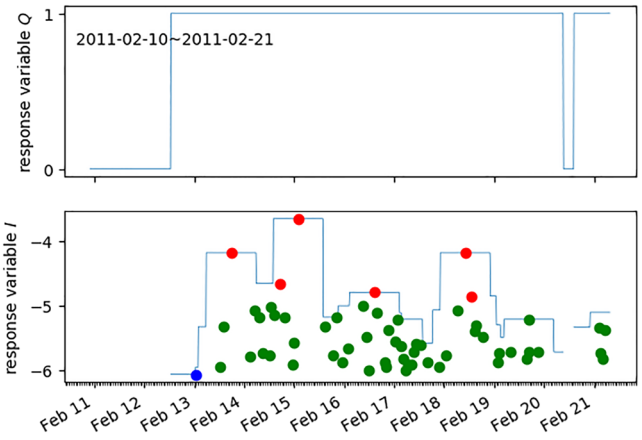

Figure 3. An example of how we define response variables based on the recorded flares that happened with HARP 377. The lower panel is the value of I given all flares, while the upper panel is the value of Q. Still, red, green, and blue points represent M/X, C, and B flares, respectively. Notice that there are missing values of I. The missing part is defined as the quiet region where correspondingly Qs take a value of 0.

Some of the HARPs have missing frames, which result in the time interval between two adjacent frames being longer than 12 min. In this case, we set up a tolerance threshold: If the number of missing frames in total for one sample input is less or equal to 10, we apply hot deck imputation (Andridge & Little, 2010) to fill the missing values. However, if there are more than 10 frames missing, we drop the sample.

2.2. Model Description

We adopt a mixed LSTM (Hochreiter & Schmidhuber, 1997) regression model to portray the relationship between SHARP parameters and flares, with a novel loss function to measure the differences between predicted results and the two-dimensional response variables defined in section 2.1.1. The LSTM model predicts outcomes using trained nonlinear transformations of input parameters and has been applied to classification of time-series data (Goodfellow et al., 2016, chapter 10). It should be noted that in Chen et al. (2019), the LSTM is only used for binary classifications, whereas in this paper, the LSTM is used for both regression and classification. We call the proposed model a mixed LSTM regression model in that it is an LSTM model combining regression and classification tasks.

2.2.1. Model Structure

The flowchart of the model is shown in Figure 5. For each sample, the input/predictor is $5m$ sets of SHARP parameters (see Figure 4), a $1 \times 5m \times p$ tensor. Again, m is the number of hours of data we use for prediction before current time point and n is number of hours from 24-hr window's left bound to now. m takes value from 6, 12, 24, and 48, which are a series of data lengths typically considered for training prediction models for solar flares; n takes values from 0, 6, 12, and 24; and p takes the value of 20, since we consider 20 SHARP parameters. The output/response is a 2×1 vector, including the predicted quiet score, $\hat{Q}$, and predicted intensity, $\hat{I}$ (see Table 4).

As shown in Figure 5, the model starts with LSTM layers. We introduce dropout layers (Srivastava et al., 2014a) between adjacent LSTM layers with dropout ratio $= 0.3$. The number of LSTM layers $= 4$, the dimensionality h of the LSTM layers and the output space is 30, and the sample size N in one batch is set to be 40. Take a model with $m = 24$ and $n = 6$ as an example. We have 38,906 samples available in training set (see section 2.1.2). For each epoch, we randomly assign them to 41,869/40 ≈973 batches. Therefore, the input is one batch out of 973, a $40 \times 120 \times 20$ tensor. After the LSTM layers, the output is a $40 \times 120 \times 30$ tensor, given $h = 30$. Then, it goes through the truncation procedure, during which the tensor becomes $40 \times k \times 30$, typically $k \ll 120$. Considering that LSTM is a sequential model for time series (Goodfellow et al., 2016, chapter 10), the choice of $k = 5m = 120$ corresponds to the sequence prediction model that explicitly adopts all these 120 input frames. However, our main goal is to capture the behavior of the $5n$ subsequent HARP frames. Therefore, the output from the latter few frames (k frames) suffice for making the desired predictions. Specifically, k takes the value of 1 in our models. Nevertheless, we have tried taking more than one ($k = 2, 5, 10...$) frames' output into the next layer and did not obtain a significantly better result.

After the LSTM and truncation layers, we feed it to two separate submodels for Q and I's training, respectively, each of which contains two dense layers. The first dense layer serves the purpose of reducing the second dimension of the tensor to 1, while the second condenses the third dimension to 1. Intuitively, the first dense layer works to combine all the information in all k frames to 1 frame for each feature and the second combines information of all p features into 1 superfeature. A `Relu` function is added between two dense layers to break the linearity. Since we take $k = 1$ in our models, the Dense Layers I_1 and II_1 shown in Figure 5 are deprecated, leaving only `Relu` functions. The only difference between these two submodels is that we further add a `Sigmoid` function at the end of the Q-training model in order to keep its value, interpreted as the probability of being unquiet, between [0,1]. Though Q and I go through two separate pipelines, they are not independent during the training. We introduce the loss function in section 2.2.2 that enables us to consider Q and I jointly in the training.

We set the epoch number to be 20. Each model takes five to seven epochs, which costs 5 to 10 min, to converge and around 20 min to finish all the 20 epochs (on a 2.3-GHz, i5, 16-GB machine that we use). Typically,

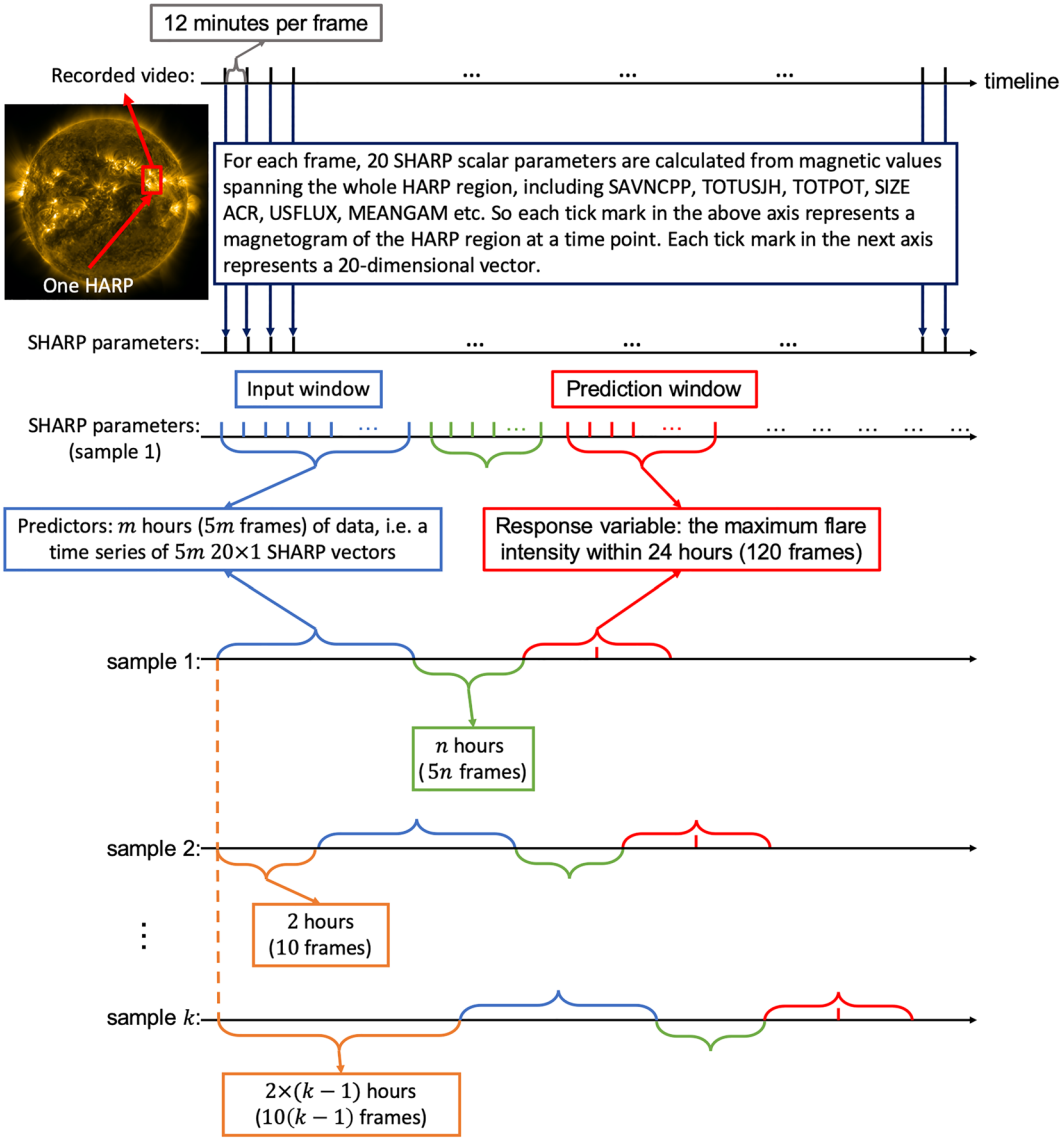

Figure 4. A diagram of how we prepare samples for training the algorithm (see section 2.1.2). For each HARP, there is a "video" containing a time series of magnetograms. For each frame, 20 SHARP parameters are calculated from the magnetic field components over the whole HARP. Therefore, we can obtain a data matrix for each HARP with 20 columns and "the number of frames (magnetograms)" rows. Data in blue braces are the predictors. Green braces denote the prediction intervals, and the response variables are decided based on the maximum flare intensities recorded in red braces. Samples are taken every 10 frames.

during the first one to three epochs, the model learns the means of all response variables and assigns the predicted intensities as the sample mean. Then, it takes a few epochs for the model to optimize over the parameters. And in the next one to three epochs, the loss converges superlinearly. Figure 6 gives a typical example of the variation of the loss function in the training process. We will give a detailed definition of the loss function in section 2.2.2.

Specifically, we here reemphasize several strategies implemented to avoid overfitting issues. First, the dropout layers with dropout ratio equal to 0.3 are set between adjacent LSTM layers. Those dropout layers randomly rule out 30% of the neurons from the preceding LSTM layers which have been proven to be an efficient way to avoid overfitting (Srivastava et al., 2014b). Second, we apply early stopping with back propagation strategy (Doan & Liong, 2004) by setting the epoch number to 20. Last and most importantly, the sample size is over 60,000–37,784 quiet samples plus 22,928 nonquiet samples after the preprocessing pipeline (section 2.1.2), which is enough for the model to learn the behavior of solar flares comprehensively.

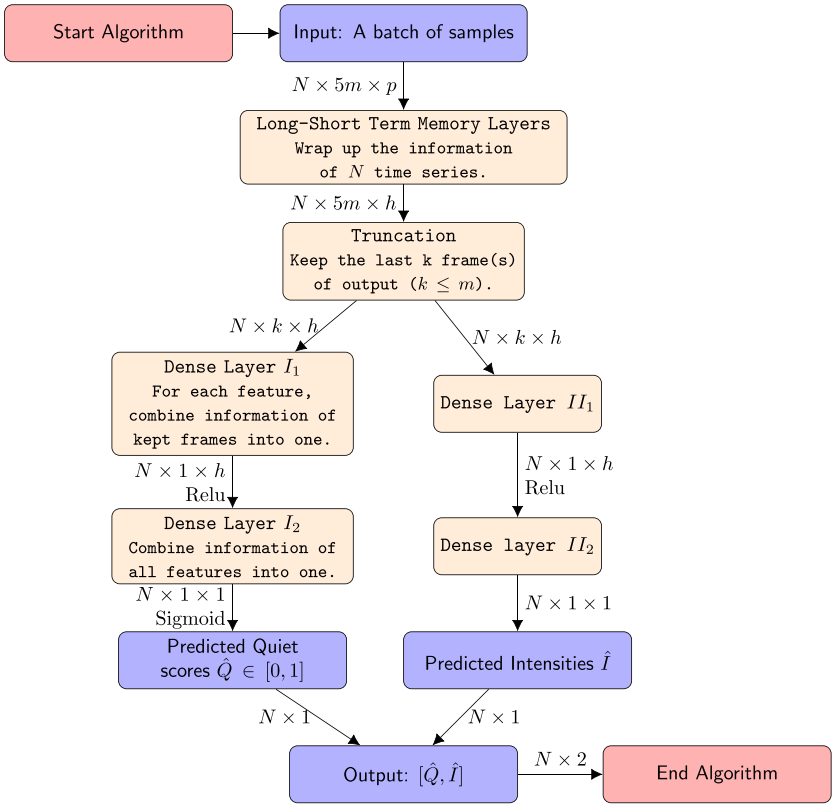

Figure 5. The flowchart of the LSTM regression model, discussed in section 2.2.1. In the figure, N is the number of samples in one batch, $5m$ is the number of frames for each sample (see Figure 4 for details), and p is the number of features we take into consideration. h is the dimensionality of the LSTM layers and the output space, and k is the number of frame(s) we keep after going through the LSTM layers.

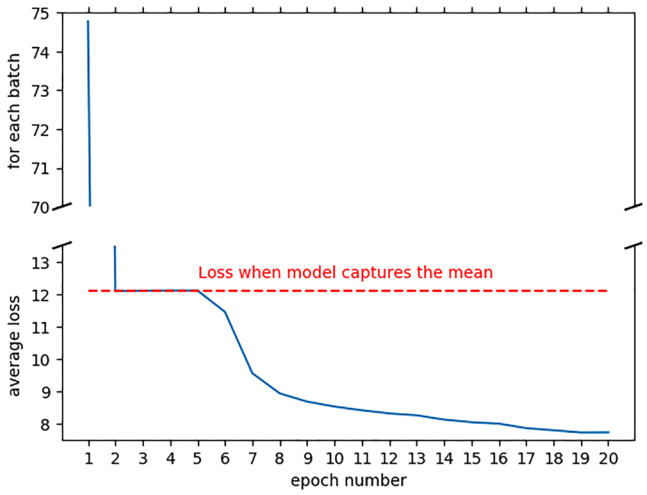

Figure 6. An example showing the convergence behavior of the mixed LSTM regression model. The x axis labels the epoch number, and the y axis stands for the average loss across batches. From Epoch 1 to Epoch 2, the average loss for each batch drops approximately from 75 to 12. In order to also visualize clearly the superlinear change starting at Epoch 5 in one figure, we cut the intermediate part of the loss change between Epochs 1 and 2.

2.2.2. Loss Function

In our mixed LSTM regression model, the response variables contain both Boolean and continuous values. Therefore, we need to adopt a special mixed approach to jointly evaluate the loss. In addition, for those samples with $Q = 0$, there are no exact values of intensity recorded. We assign N/A to those "missing" intensity values. The desired loss function should avoid the usage of I for those samples with intensity values missing. We use binary cross-entropy loss in terms of $\hat{Q}$, which takes values between 0 and 1, and the squared error loss for $\hat{I}$ (Janocha & Czarnecki, 2017), which takes values in $\mathbb{R}$; see Table 5 for examples. Furthermore, we define three tuning parameters to flexibly deal with the overabundance of the quiet samples and the noncomparability between the loss for quiet score and that for (logarithm) intensity values.

More precisely, the loss function for each batch is defined as

$$\mathscr{L} = \sum_{\text{batch samples}}^{N} \begin{bmatrix} r & 1 \end{bmatrix} \begin{bmatrix} -\log(1-\hat{Q}) - \log(\hat{Q}) 0 (I-\hat{I})^2 \end{bmatrix} \begin{bmatrix} 1(Q=0)w_1 \\ 1(Q\neq 0)w_2(I) \end{bmatrix}$$

$$= \sum_{\text{batch samples}}^{N} \left[-1(Q=0)w_1 r\log(1-\hat{Q}) + 1(Q\neq 0)w_2(I)(-r\log\hat{Q} + (I-\hat{I})^2) \right],$$

where Q only takes values in the binary set {0,1}, $I \in [-7, -3]$ are observed log-intensity values, $\hat{Q} \in [0, 1]$ and $\hat{I} \in \mathbb{R}$ are fitted values, $1(Q = 0)$ is the indicator function for $Q = 0$, and N is the sample size of

Table 5
We Use Binary Cross-Entropy Loss in Terms of $\hat{Q}$ and L_2 Loss for $\hat{I}$

Loss	Quiet sample	Nonquiet sample
Q	$-\log(1-\hat{Q})$	$-\log(\hat{Q})$
I	N/A	$(I-\hat{I})^2$

each batch. We take $N = 40$ in all our models (see section 2.2.1). The tuning parameters w_1, $w_2(\cdot)$ and r are adopted to calibrate the weight of each component in the loss function. Specifically, w_1 is the weight for loss generated by quiet samples, while $w_2(.)$ is a function set for nonquiet samples returning weights given specific intensity, and r is the weight for the loss generated by the Q dimension. Note that for the loss function, only the relative values of w_1,$w_2(\cdot)$ and r matter—a loss function can be defined up to a positive constant. Next, we explain the different components in the design of this loss function.

For the loss generated by the Q dimension, since $Q \in \{0,1\}$ and $I \in [-7, -3]$, the scale of Q's loss is incomparable to I's loss. We multiply the Q dimension's loss by a scale parameter r for all samples in order to balance the losses of Q and I. In terms of loss of quiet samples, there are significantly more of them, 37,784, than nonquiet samples (flare events), 22,928. We note that our main focus is on those nonquiet samples when predicting local maximum flare intensities. Therefore, we multiply the loss of the quiet samples with weight $w_1(<1)$ in order to attenuate the impact caused by the overabundance of quiet samples when training our prediction models. The values of r and w_1 are both tuned by the cross-validation (Hastie et al., 2009, chapter 7.10). Specifically, we consider r taking values in set {1, 2, 5, 10, 15} and w_1 taking values in set {0.1,0.2,0.5,1}. We randomly divide the training data set into 10 folds. For each possible pair of r and w_1, we train the model 10 times with nine folds as the training set and the remaining fold as the testing set. Finally, we take the parameter values $r = 5$,$w_1 = 0.2$, which results in the lowest average loss.

Now we consider the loss associated with the nonquiet samples (flare events). As we can see in Figure 2, C flares dominate the data set while the samples for B and M/X flares are comparatively more limited. We adopt the squared error loss for the prediction of flare intensities. If we simply weight all the input samples equally, under the square loss setting, the consequence is that the predicted results will tend to cluster at the central part (around -6 to -5.5 for logarithm intensity, corresponding to C flares), which are the 30% and 70% quantiles of the response variables, respectively, instead of the $[-7, -3]$ intensity range. This is inconsistent with our original intention that M/X flares shall stand out from other flares as much as possible in the model. Thus, we add $w_2(\cdot)$ (see equation (1)), which serves to balance the weights of samples from different classes, which downweight the prevalent C flares essentially. We define the weight for the flare with intensity level I as

$$w_2(I) = |I - \mu| \times \text{constant}. \tag{1}$$

Next, we explain our rationale for choosing this particular set of weights. We fit the empirical distribution of the logarithm of the flare intensity of the full data set to a Cauchy distribution, which is a heavy-tailed distribution, with location parameter $\mu = -5.84$ and scale parameter $\gamma = 0.31$. The fitted curve is shown in Figure 2. The weight is set to be the L_1 distance from μ multiplied by a constant specified based on the proportion of the quiet samples. By doing so, we maintain the balance of samples of M/X, C, and B classes. Equation (2) gives the detailed probability mass corresponding to each flare class under the weighting scheme given by equation (1):

$$\begin{cases} \text{B flares: } \int_{-7}^{-6} |x-\mu| \cdot f(x)dx = 0.121 \\ \text{C flares: } \int_{-6}^{-5} |x-\mu| \cdot f(x)dx = 0.116 \\ \text{M/X flares: } \int_{-5}^{-3} |x-\mu| \cdot f(x)dx = 0.114 \end{cases}, \tag{2}$$

where a Cauchy distribution with location parameter μ and scale parameter γ has probability density function denoted by $f(x) = \left[\pi\gamma\left(1+\left(\frac{(x-\mu)}{\gamma}\right)^2\right)\right]^{-1}$.

With this strategy, we can combine the quiet and nonquiet samples in one model and train them simultaneously. Again, the loss function $\mathscr{L}$ is defined over each batch with N samples therein. Therefore, we can obtain the "number of batch" of losses for each epoch. The loss we evaluate and visualize in Figure 6 is the average loss of all batches over each epoch. The results calculated based on the loss function $\mathscr{L}$ are shown in section 3.1.

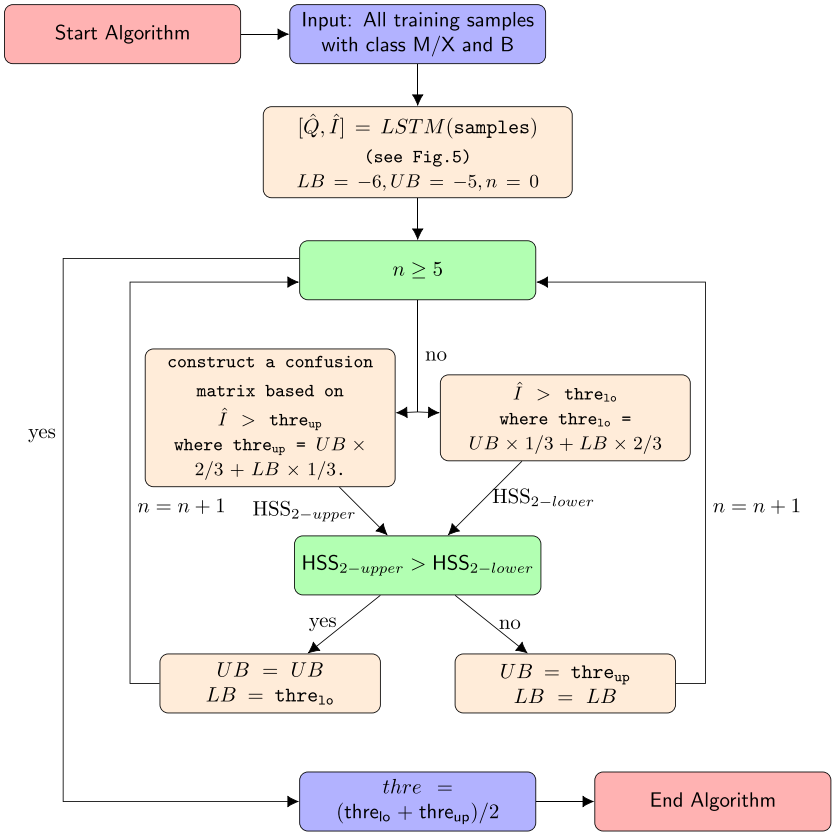

Figure 7. The flowchart of M/X versus B classification, discussed in section 2.3. After inputting all training samples with class M/X and B into the trained LSTM model, we use the output $\hat{I}$ together with I to decide an optimal threshold between M/X and B with trisection method. The loop time is set to 5.

2.3. Extension to Classification Models

In this section, we introduce *binary classification models* that are built upon the mixed LSTM regression model in section 2.2. The binary classification models are designed for classifications of M/X versus B, M/X versus B/Q, and M/X versus C/B/Q.

For M/X versus B, that is, strong/weak flare classification, we only consider training samples that have flare intensities ranging from $[-7, -6) \cup [-5, -3)$. Borrowing the idea from transfer learning in Yosinski et al. (2014), we make use of the output given by the mixed LSTM regression model, $\hat{I}$, to decide an optimal threshold between M/X and B flares.

Since we know the observed intensity, I, of all training samples, for each potential threshold (thre $\in (-6, -5]$) for $\hat{I}$, we can construct a confusion matrix, where true positives TP $= \sum 1(\hat{I}_i \geq \text{thre}, I_i \geq -5.5)$, false positives FP $= \sum 1(\hat{I}_i \geq \text{thre}, I_i < -5.5)$, false negatives FN $= \sum 1(\hat{I}_i < \text{thre}, I_i \geq -5.5)$, and true negatives TN $= \sum 1(\hat{I}_i < \text{thre}, I_i < -5.5)$, where each term is summed over all available training samples. Then we can calculate the HSS_2 score correspondingly (see Bobra & Couvidat, 2015 for the definition of HSS_2). Again, $1(\cdot)$ is an indicator function. Note that, in this case, I only takes values in $[-7, -6) \cup [-5, -3)$. Any number between -6 and -5 could act as the threshold for observed intensity, I. We hereby take the value of -5.5.

Next, we apply the trisection method (Gu et al., 2006) to find the threshold that yields the highest HSS_2. For each iteration, we obtain a thre_{lo} and a thre_{up} by trisecting the current range of threshold. By constructing confusion matrixes respectively, we compare the HSS_2 score, choose the one with the higher score, and define new thre_{lo} and thre_{up}. Throughout the iterations, the range of possible thresholds keeps getting smaller, and finally, we reach an optimal threshold for $\hat{I}$. The flowchart of the algorithm is in Figure 7.

The M/X versus B/Q classification model adopts the same strategy as the M/X versus B classification model does on determining the threshold between M/X and B/Q. Different from the M/X versus B/Q and M/X

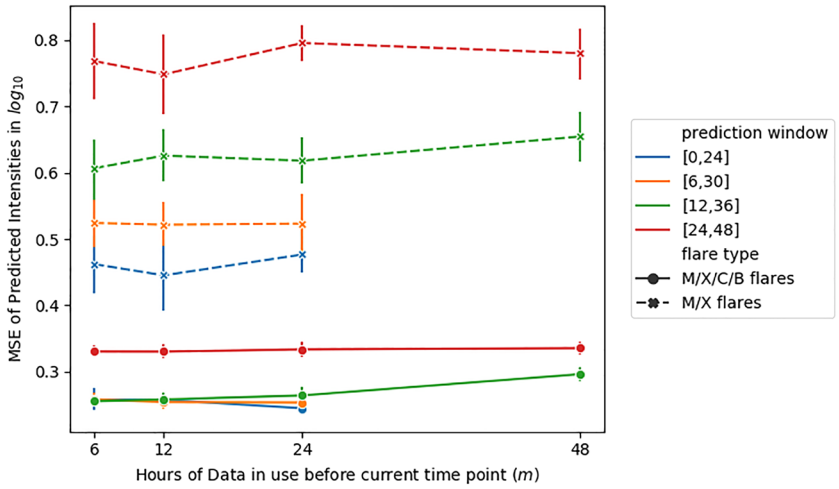

Figure 8. Line chart showing the MSEs of all mixed LSTM regression models, shown in section 3.1. Again, m is the number of hours of data we use before the current time point, $[-m,0]$ is the input window, and $[n, n + 24]$ is the prediction window. Each point with a vertical line is the average MSE and its 95% confidence interval of 10 regression models with the same $[-m,0] - [n, n + 24]$ trained separately. Each line shows the variation of MSE for models with the same prediction window and different lengths of input windows. The solid lines represent the MSEs of all nonquiet testing samples (M/X/C/B). The dashed lines represent the MSEs of those testing samples with M/X flare intensities.

versus B models, the M/X versus C/B/Q classification model no longer has the sweet $[-6, -5)$ buffering area for us to train a threshold. Once we include C flares in the model, the threshold is fixed at -5.

We use the following six metrics to evaluate all our binary classifiers: recall, precision, the F_1 score, the Heidke skill scores (HSS_1 and HSS_2) (see Bobra & Couvidat, 2015 for the definition of HSS_1 and HSS_2), and the true skill statistics (TSS), among which HSS_2 and TSS are our main focuses. Specifically, recall and precision are two standard metrics evaluating the quality of a prediction. The F_1 score is the harmonic mean of recall and precision. However, these three scores can be rather unstable when encountering unbalanced samples, which is true in our case where the B/C flares outnumber the M/X flares. We consider TSS and HSS_2 as two reasonable measures of classification performance for solar flares. TSS is invariant to the frequency of samples, unlike recall or precision. HSS_2 measures the fractional improvement of the forecast over the random forecast. There are detailed descriptions of HSS_1, HSS_2, and TSS in Florios et al. (2018). Bloomfield et al. (2012) give conceptual comparison and discussion on the suitability of these metrics when predicting solar flares. A summary of the binary classification results is shown in section 3.2.

2.4. Test Samples Preparation

In this paper, we adopt the following strategy for preparing the testing samples to give a fair evaluation of the performance of our algorithms. Recall that each sample is a time series of SHARP parameters and corresponds to a two-dimensional response variable $[Q, I]$.

First, we take all the samples from the full data set after 2015 (see how we get full data set and do training/testing splitting in section 2.1). For each sample with corresponding response variable $Q = 1$ (nonquiet samples), there should be at least one flare happening in the 24-hr time window and the maximum intensity of all the applicable flares should be equal to I. For samples with overlapping predictors and the corresponding response variables belonging to the same flare class, we keep one of them at random to avoid repeated predictors—response variable pairs in the testing set. Quiet samples are collected with the same strategy. Section 3 and Appendices A, B, and C give results for using testing samples obtained via this strategy.

3. Results

In this section, we present results in sections 3.1,3.2, and 3.3 based on the models described in section 2. In section 3.4, we illustrate that under the LSTM architecture, the most efficient time range for predicting the

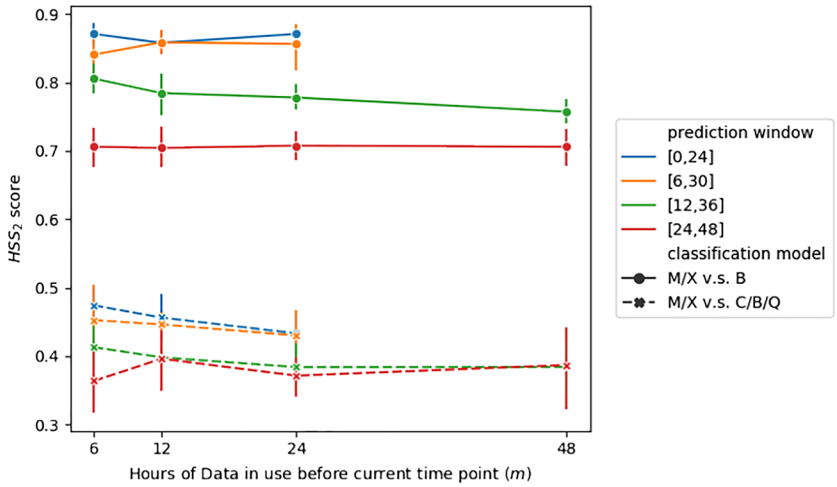

Figure 9. Line chart showing the HSS_2 scores of all classification models, covered in section 3.2. Similar to Figure 8, each point with a vertical line is the average HSS_2 and its 95% confidence interval of 10 classification models with same $[-m,0] - [n, n + 24]$ trained separately. Each line shows the variation of HSS_2 for models with the same prediction window and different lengths of input windows. The solid lines represent the HSS_2s of M/X versus B models. The dashed lines represent the HSS_2 scores of M/X versus C/B/Q models.

solar activity using the SHARP parameters is within 24 hr before the prediction time. Finally, case studies of intensity prediction with several representative HARPs are given in section 3.5.

With the current time point specified as time 0, we denote a model as "$[-m,0] - [n,n + 24]$" if it uses data in time range $[-m,0]$ to predict maximum local flare intensities within the $[n,n + 24]$ time window ($n,m \geq 0$). We define the $[n,n + 24]$ time window as *prediction window* and $[-m,0]$ time window as *input window*. For example, if we want to use the past 6 hr of data to predict the maximum local flare intensity in the 24-hr window [0,24], the model is denoted as $[-6,0] - [0,24]$. The prediction window is [0,24] and the input window is $[-6,0]$ in this case. Similarly, if we want to use the past 12 hr of data to predict the maximum local flare intensity in the next [12,36] hours, the model should be denoted as $[-12,0] - [12,36]$. The prediction window is [12,36] and the input window is $[-12,0]$.

To allow fair comparisons across models, models with the same prediction window but different input windows are applied to the same group of samples. Consider a series of models $[-6,0] - [0,24]$, $[-12,0] - [0,24]$, $[-24,0] - [0,24]$ as an example. Their samples are all filtered based on the standard for model $[-24,0] - [0,24]$ (see section 2.1.2 for details on sample preparation). Therefore, for each sample, we have 24-hr length of SHARP parameters as the predictors, while we only use the last 6 and 12 hr of predictors for models $[-6,0] - [0,24]$ and $[-12,0] - [0,24]$.

3.1. The MSEs From the Mixed LSTM Regression Model

In this section, we present the MSEs of predicted $\log_{10}$ flare intensities from all models in the of line charts. The complete MSE tables for all models and all classes of flares can be found in Appendix A.

Figure 8 is a line chart showing the MSEs for models with the same prediction window as the length of input window (m) increases (solid lines). The chart also includes the MSEs of the samples with M/X flares (dashed lines). As the prediction window gets farther away from the current time point (n increases), the MSE of all flare samples does not change too much. However, this is not true when we look at MSE calculated from M/X flares only. This shows the sensitivity of the evaluation metric, MSE, with respect to the samples that we use to calculate with. Therefore, the MSE of M/X flares can be considered as another metric for evaluating the performance of the regression models.

Intuitively, the smaller the n, that is, the closer the prediction window from the current time point, the smaller the MSE will be. This is confirmed in Figure 8. Generally, from the results, the MSE is kept under 0.3 when the prediction window is [0,24], [6,30], or [12,36]. We can keep the MSE of M/X flares under 0.5 when $n = 0$, that is, prediction window is [0,24]. We also observe that there is a sudden increase in terms of the

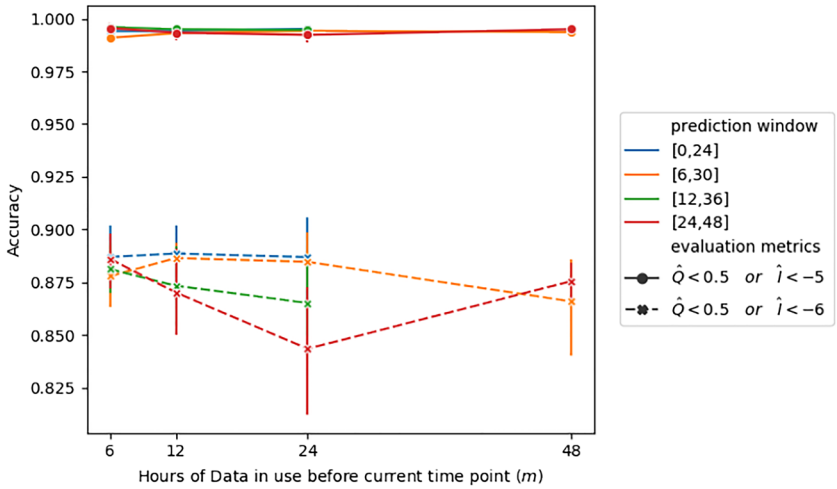

Figure 10. Line chart showing the classification accuracy of quiet samples in all models, covered in section 3.1. Each point with a vertical line is the average accuracy and its 95% confidence interval of 10 models with same $[-m,0]-[n,n+24]$ trained separately. Each line shows the variation of the accuracy for models with same prediction window and different lengths of input windows. The solid lines represent the accuracy when the evaluation metric is $\hat{Q}<0.5$ or $\hat{I}<6$. The dashed lines represent the accuracy when the evaluation metric is $\hat{Q}<0.5$ or $\hat{I}<5$.

MSE of M/X flares when the prediction window is shifted from [6,30] to [12,36] and [24,48]. However, we do not observe any significant patterns of the MSE varying monotonically as a function of m, the length of the time series that we use for prediction. We elaborate discussions on these results in section 3.4.

3.2. Performance of the Classification Models

We use the HSS_2 score to compare the performances of M/X versus B and M/X versus C/B/Q classifiers. Results in other metrics mentioned in section 2.3 are shown in Appendix B. In addition, since M/X versus B/Q models give us similar HSS_2 scores as M/X versus B models do, we also put results of M/X versus B/Q models in Appendix B.

The HSS_2 score results are also shown in the form of a line chart in Figure 9. There is a large gap between all M/X versus B models and all M/X versus C/B/Q models. As mentioned in section 2.3, we have an intensity interval, $[-6,-5)$ (for C flares) where there is no flare defined as M/X or B. This is mainly why we can get incredibly high scores ($HSS_2>0.8$ when the prediction window is [0,24] or [6,30], $HSS_2>0.7$ when all models) for M/X versus B. As for the M/X versus C/B/Q model, we can hardly get HSS_2 scores greater than 0.5. We manage to classify roughly half of the M and X flares out of other flares when prediction window is [0,24] (see Appendix B). Almost all of the misclassified M and X flares have predicted intensities falling into C flares' intensity range (see Figure 13). We do not observe an obvious HSS_2 score difference between models with prediction window [0,24] and [6,30]. But when the prediction window is shifted from [6,30] to [12,36] and [24,48], there is a large decrease in terms of the HSS_2 score.

3.3. Results of Quiet Samples From the Mixed LSTM Regression Model

In sections 3.1 and 3.2, we only summarize the prediction results of nonquiet samples, that is, samples with response variables $Q=1$. In this section, we will particularly focus on the performance of all the models in terms of the quiet samples, that is, samples with response variables $Q=0$.

First, we examine the fitted distribution of the predicted intensity ($\hat{I}$) of the quiet samples in Figure 13. This is an example of a $[-6,0]-[0,24]$ model. We observe that almost all of the quiet samples have $\hat{I}<-5$ in the testing set, which indicates that the false alarm (false positive rate) of quiet samples can be restrained significantly in our models. Next, we formally evaluate the performance of the prediction. Note that we don't have the exact observed intensity ($I=N/A$) for quiet samples (see examples of how we define response variables in Table 4). Therefore, we consider the prediction result ($[\hat{Q},\hat{I}]$) as successful if it meets either of the following two requirements: (1) the predicted intensity $\hat{I}<k$ and (2) predicted quiet score $\hat{Q}<0.5$.

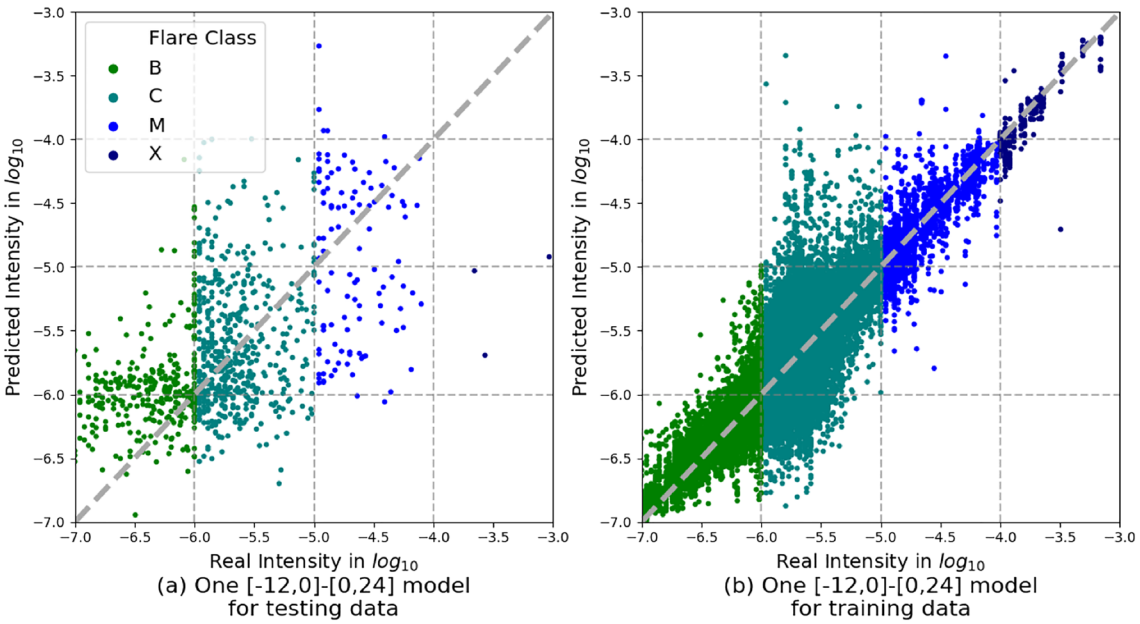

Figure 11. Predicted intensities versus true intensities. Each point represents a recorded flare. Purple stands for X flare, blue for M, aqua for C, and green for B. For both panels (a) and (b), the x axis is the observed intensity, and y axis is the predicted intensity. The thick gray dashed line $y = x$ shows the ideal positions where every point should locate when being accurately predicted.

Specifically, k takes the values of -5 and -6, where $k = -5$ evaluates the rate of falsely predicting a quiet sample as intensive flare (M and X flare) while $k = -6$ evaluates the rate of falsely predicting a quiet sample as M, X, or C flare. We denote $k = -5$ as Metric 1 and $k = -6$ as Metric 2.

Figure 10 shows the summarized result of the quiet sample prediction, where the solid line corresponds to Metric 1 and the dashed line to Metric 2 (the summary table can be seen in Appendix C). We obtain an accuracy of over 98.5% for all models in terms of Metric 1 and over 80% in terms of Metric 2. Recall that "-5" is the cutoff of the logarithm of flare intensity for B and C flares; thus, as long as we don't give a $\hat{I} > -5$ which is an alarm of intense flare, we can consider the prediction satisfying. Therefore, we conclude that our regression models have an excellent performance on restraining false alarms.

3.4. Post Hoc Analysis

In this section, we show visualizations of the prediction results, combined with the regression and classification results shown in sections 3.1,3.2, and 3.3, to investigate in-depth how the information in the data (time series of SHARP parameters) convey for solar flare predictions under the LSTM architecture.

Figures 11 and 12 show the predicted intensity against the observed intensity with each point representing a flare event. Each color in the figures represents one class of solar flare. Purple stands for X flare, blue for M, aqua for C, and green for B. Specifically, except that Figure 11b is plotted based on the training samples, all other subpanels in Figures 11 and 12 are plotted based on testing samples corresponding to five models with different prediction windows and input windows. Figure 11b exhibits the best performance over all figures, since it is based on a training set. We cannot expect to achieve this high accuracy when applying models to the testing set.

Figure 13 shows the fitted Gaussian distribution of each class's predicted intensity. The left panel is the fitted Gaussian distribution for training samples, and the right panel is for testing samples. Each color represents one class of flares. It can be seen that the different classes of flares, especially neighboring ones, have overlapping predicted intensity values. Nevertheless, the strong flares and weak flares (or quiet time) are still highly distinctive.

Not surprisingly, the farther the prediction window from the current time point, the worse the prediction results. This is also intuitive: Predicting what happens after 1 hr is easier than predicting what happens

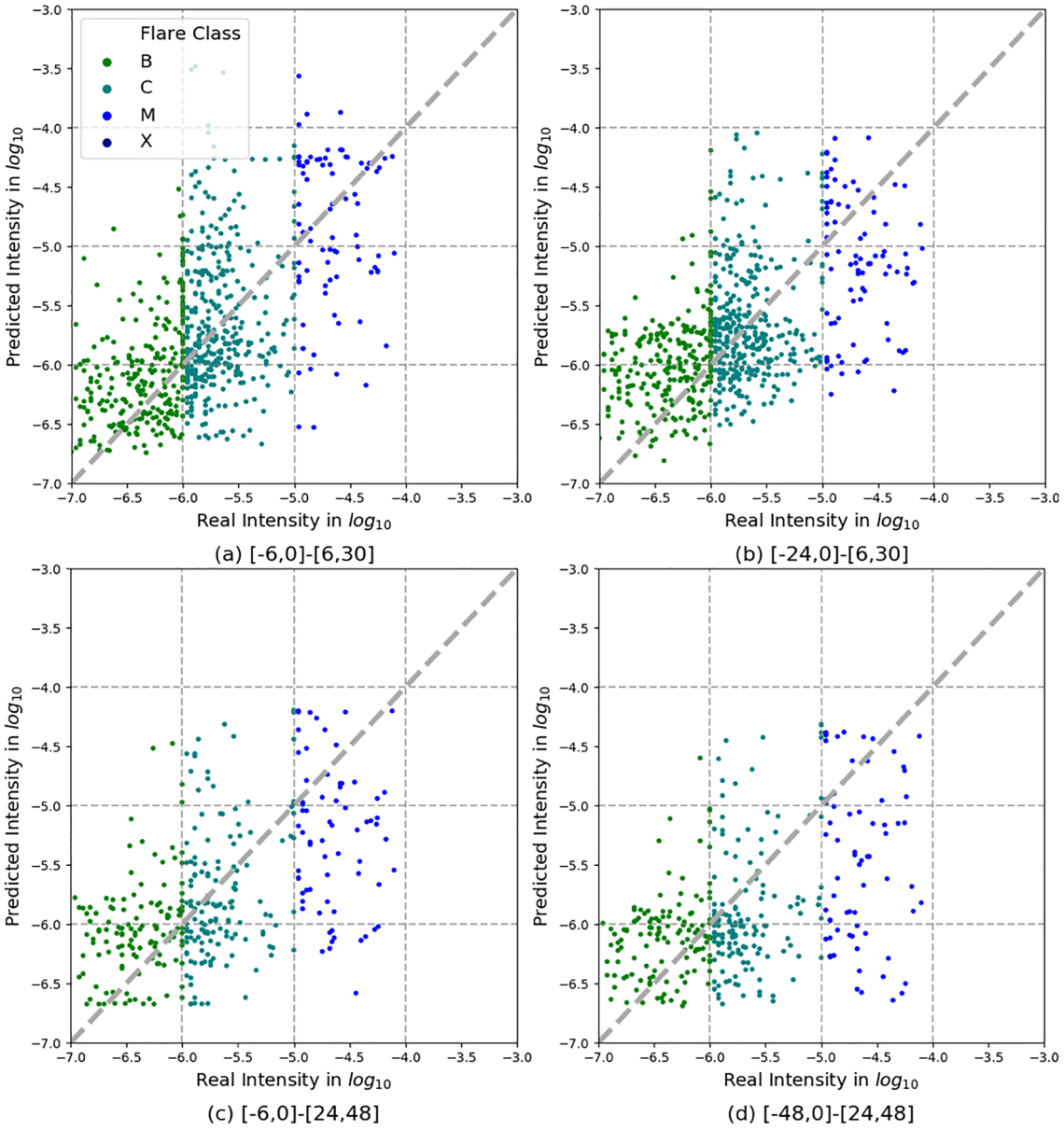

Figure 12. (a–d) Visualizations for four example models. The figures share the same setting as Figure 11. Note that, in both Figures 12 and 13, there is no X flare plotted. Recall that we define the prediction window as [n,n+24]. Generally, there are no applicable X flares in testing set for $n > 0$. We have very few X flares. Most of them happened before 2015. For the limited X flares that happened after 2015, they either have many frames missing before it happened or happened only a few hours after the video starting. So we don't have X flares in testing set for models with prediction windows farther away from the current time point.

after 10 hr. Another finding is that considering more data backwards (greater m) does not necessarily guarantee a better prediction result. The explanation is twofold.

First, we speculate that the most useful information for predicting the behavior of the prediction window is within 24 hr beforehand. Here 24 denotes the hours from the center of the prediction window to now. Once $n+12 \geq 24$ (12 is half of the prediction window's length), considering more information does not help much based on our results. Notice that, even though the TSS and HSS_2 scores decrease as the n increases, they always experience a sharp drop when the prediction windows move farther away from [6,30] to [12,36]; that is, n increases from 6 to 12 in all models. Recall that k in Figure 5 is the number of frame(s) we kept after going through LSTM layers, and we take $k = 1$ for all our models. Therefore, we are essentially using the output information of the last frame (n hours from the prediction window) to predict the behavior in the

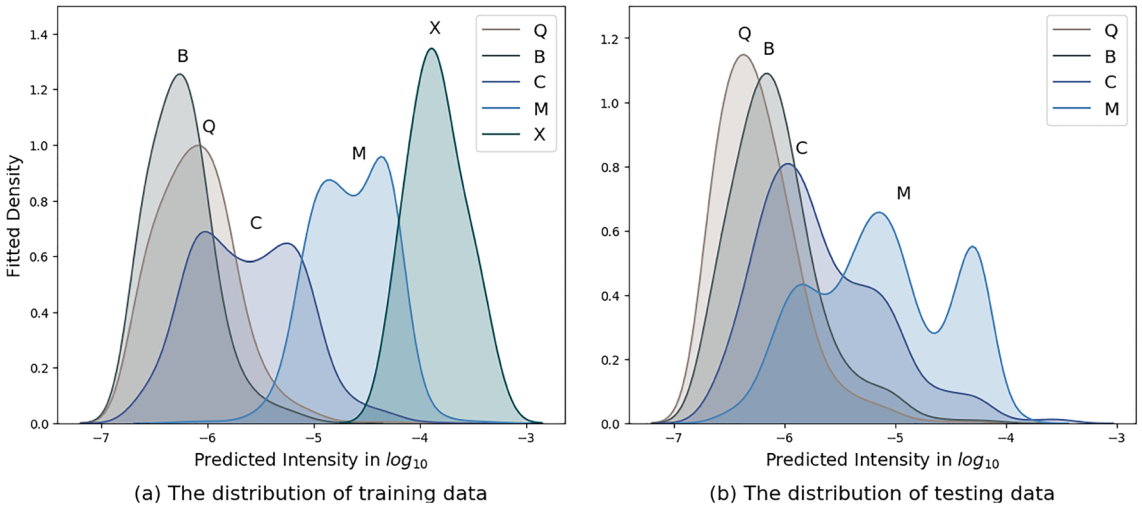

Figure 13. Fitted distribution of predicted intensities based on one $[-6,0] - [0,24]$ model. The distribution is fitted using Gaussian kernel with bandwidth $= 0.15$. The x axis plots the values taken by predicted intensities, and the y axis stands for the density of fitted distribution. Ideally, flares with class B, C, or M should follow an asymptotically normal distribution. The predicted distribution (a) for training data is close to the ideal setting, while for the testing set (b), the predicted intensities are still having a hard time separating themselves from other flares.

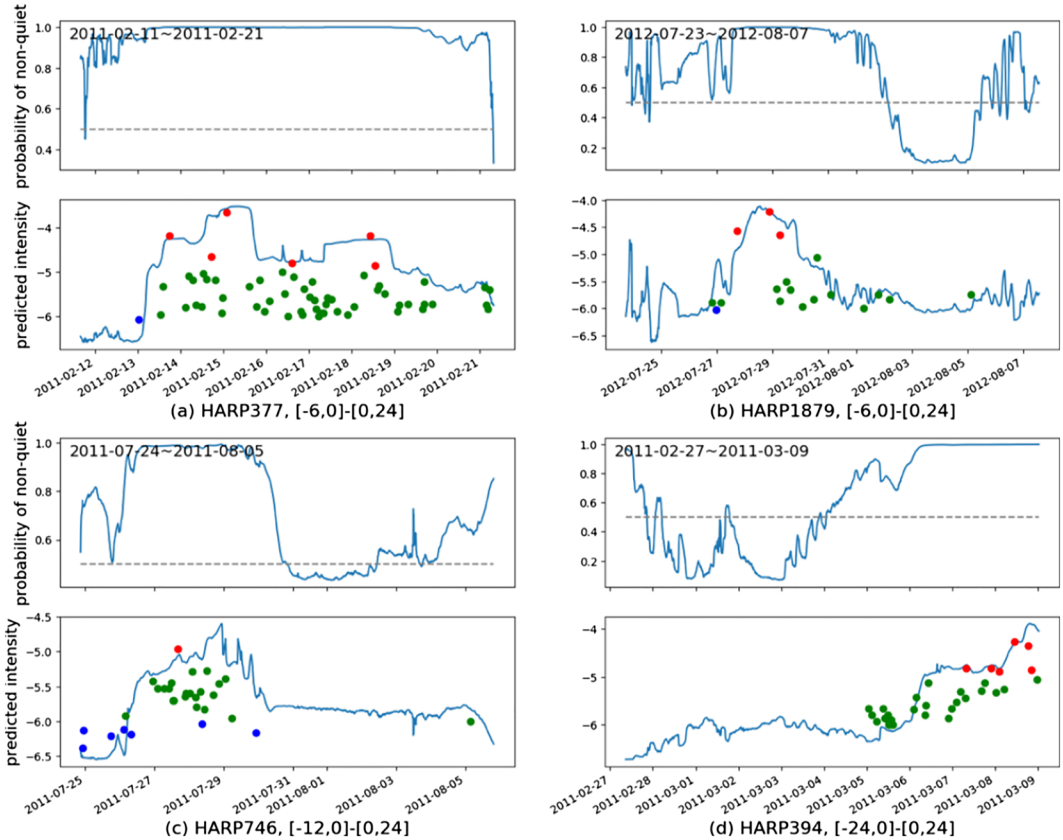

Figure 14. (a–d) Case studies: successful cases. For each plot, the blue curve on the upper panel is the predicted $\hat{Q}$ score. The gray dashed line taking the value of 0.5 is the threshold of dividing quiet and nonquiet times. The blue curve on the lower panel is the predicted real-time flare intensity, $\hat{I}$. There is no time shift on each plot. Each red, green, or blue round point corresponds to one recorded M/X, C, or B flare, respectively. Unlike in Figure 1, the height of each point is exactly the $\log_{10}$ intensity of the flare it represents.

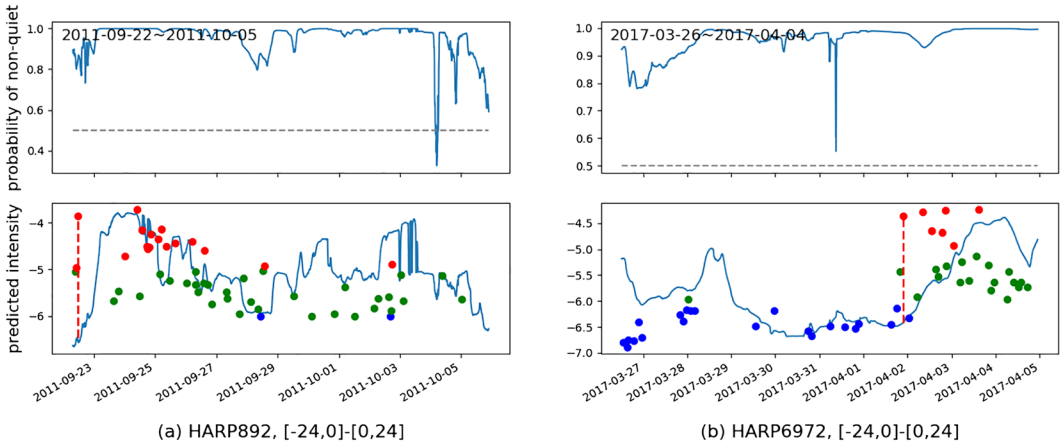

Figure 15. (a,b) Case studies: failed cases. Same setting as Figure 14. In addition, the red vertical dashed line is to indicate the largest prediction error.

prediction window. A worse result indicates that the last frame is less relevant to the prediction window or it is harder for LSTM to build a relationship between the prediction window and the last frame. Thus, the sharp drop when the prediction window shifts from [6,30] to [12,36] indicates that the solar activities within the 24-hr window prior to the events have a significant influence on the behavior in the prediction window.

Second, even though the most useful information for prediction is within 24 hr before the events, considering more information offering us worse result is still counterintuitive. This is due to the limitations of the LSTM model. The LSTM is an artificial recurrent neural network (RNN) architecture used for digging out the temporal properties within time-series data. The parameter matrices for each gate remain unchanged for all input time series. Therefore, the LSTM considers the entire time evolution process in a homogeneous way. If the whole time series before the event is not acting homogeneously, adding information 24 hr before can, on the contrary, impair the performance of the prediction.

3.5. Case Study

In the case study section, we focus on the model performances on M and X flares' predictions for two reasons. First, M and X flares are of primary concern in the flare prediction problem. Second, as shown in Figure 8, the model can already offer us a decent prediction, that is, a relatively small MSE, for B and C flares. Besides, Figure 13 shows that, for both the training and testing sets, quiet samples' predicted intensities are restricted below -5. Hence, M and X flares are not only the most important but also the most difficult flares to predict, that is, generating the highest MSE.

Figures 14 and 15 show six prediction plots, including four well and two badly performed examples, each of which corresponds to one HARP and one model. The four well-performed examples in Figure 14 are chosen where at least one of their M and X flares lays near the $y = x$ diagonal line in Figures 12a and 12b. For the two badly performed cases in Figure 15, we choose two videos where one of their M or X flares has the largest prediction error ($|I - \hat{I}|$) among all M and X flares in the training set and testing set, respectively, in a $[-24,0] - [0,24]$ model.

A successful case should have the blue curve in the lower panel of each plot locating as close as possible to the local maximum flare, that is, the local highest round point. Note that the existence of dimension Q in the response variable is only to compensate for the nonobservable flares. Thus, the quiet score $\hat{Q}$ in the upper panel is more than a signal instead of an exact prediction result. As long as the lower panel offers a $\hat{I} \leq -6$, we can still consider the model as having a good prediction of the quiet time.

The two cases shown in Figure 15 represent two typical situations where M and X are wrongly predicted. (1) The model does perceive the increase in flare intensity but not precisely, like in Figure 15a. Predicted intensity may have increased hours before or after the intensive flares' happening. (2) The model fails to detect the intense flares totally, like in Figure 15b. However, this scenario only happens when certain M/X flares lay at the head or tail of the video. Moreover, videos also tend to have a few frames missing at

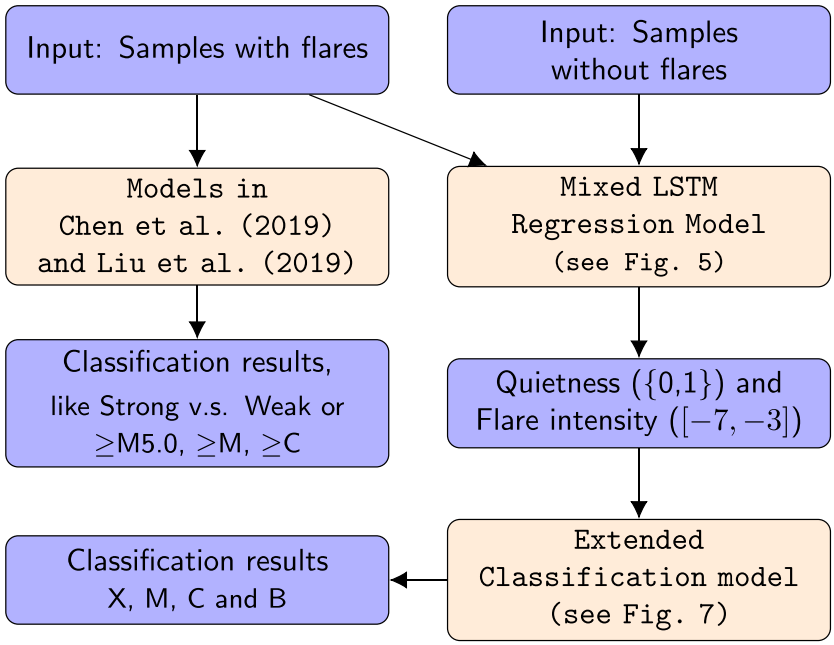

Figure 16. A direct comparison between models introduced in this paper and models used in Chen et al. (2019) and Liu et al. (2019). Mixed LSTM models can accept any kind of sample inputs and give a more informative prediction including the quietness (quiet or unquiet), flare intensity, and flare class within a 24-hr window. As a contrast, the models in Chen et al. (2019) and Liu et al. (2019) can only give classification results.

the beginning and the end. Thus, we speculate that it is the potential problem of the missing frames and the mismatch of HARP and active regions (see section 2.1.1 for details) rather than the model that restricts the performance of the prediction. We also note that there are many missing B and C flares in the GOES data set, which might reduce precision of the response variable, leading to biased prediction results.

4. Summary and Discussion

In this paper, we presented a pipeline to prepare and analyze data from the SHARP parameters and GOES data set. A mixed LSTM regression model was introduced and applied, and we shared encouraging results on solar flare intensity prediction and classification. The work in this article can be considered as one further step from the papers discussing flare classification, including Chen et al. (2019) and Liu et al. (2019).

We refer models in this paper as model$_A$ and models in the above two papers as model$_B$. Generally, model$_A$ differ from model$_B$ in several aspects (a direct comparison on the breadth of usage between models is shown in Figure 16).

- Model$_A$ consider the intensity of each flare as a continuous variable on the $\log_{10}$ scale, ranging from [−7, −3], instead of a single label, defined as a binary (strong and weak) or multiclass (≥M5.0, ≥M, and ≥C class) label. Therefore, model$_A$ could predict both the intensity and the class of the flare as opposed to only predicting flare class in model$_B$. Model$_A$ are regression models, whereas model$_B$ are classification models. For example, we consider two flares with intensity level M1.0 and C9.9. These two flares are similar in the regression model since their $\log_{10}$ intensities are close to each other but are totally different in classification model since the former is an M-class flare and the latter is a C-class flare.
- In model$_A$, we assign each frame the maximum flare intensity of flares happened within a 24-hr time window (12 hr before and 12 hr after). By doing so, model$_A$ can assign every frame a flare intensity, including the frames where there is no flare happening.
- In our notation, a time point (one frame) with no flare happening includes two cases: (1) There exists at least one flare within the 24-hr window but not at the exact time and (2) there is no flare within the 24-hr window. We consider the latter frames as quiet regions and the former together with frames with flare

happening as unquiet regions. Hence, model$_A$ can predict the quietness of a 24-hr window, instead of presuming there is flare happening at the prediction time point and classifying flare labels like model$_B$ do.
- The extended classification model in model$_A$ is only a by-product of the regression model. The way we get the classification relies on the predicted numerical flare intensity values and the trained thresholds. This is also different from the classification methods in model$_B$.

Specifically, compared to our previous results in Chen et al. (2019), the models presented in this paper stand out in several aspects.

- The prediction score, TSS and HSS$_2$ of M/X versus B, is increased by 0.1 when the prediction window is [0,24].
- We consider more cases, including [−6,0] − [0,24], [−12,0] − [0,24], [−24,0] − [0,24]; [−6,0] − [6,30], [−12,0] − [6,30], [−24,0] − [6,30]; [−6,0] − [12,36], [−12,0] − [12,36], [−24,0] −[12,36], [−48,0] − [12,36]; [− 6,0] − [24,48], [−12,0] − [24,48], [−24,0] − [24,48], [− 48,0] − [24,48], and prepare the data to offer fair comparison with same prediction windows.

There are several promising areas for future work. First, as we mentioned at the beginning of section 2.1.1, there exists a potential mismatch of the SHARPs and GOES data, which may cause bias for prediction models. We plan to address this problem in future work using flare location data. Second, the Sun's activity level experiences an 11-year cycle, where the 24th cycle began in December 2008 (Space Weather Prediction Center, 2019). The boundary between the training and testing sets in this paper are set at year 2015. Flares events that happened after 2015 are not exactly equivalent or comparable to flares before 2015. It would be worthwhile to explore other splits of the data sets into training and testing subsets. Third, in our models, we consider videos of different HARPs equally, which is certainly not the case due to the intrinsic variability among different HARPs. Moreover, there is a latent dependency among flares in the same HARP, which are not modeled in our LSTM approach. Lastly, as mentioned in section 3.5, our results are limited by its sole dependency on the SHARP parameters, which may or may not fully capture the information of the magnetic field. In the future, we plan to directly work with the HMI magnetograms for real-time prediction of flares.

Appendix A: MSE Table for Mixed LSTM Regression

In this table (Table A1) and all the following tables in the appendix, we denote the [$-m$,0] $-[n,n+24]$ model as $(n+12)-m$ for simplicity. For example, [−12,0] − [0,24] is 12-12 and [−24,0] − [24,48] is 36-24. Note that the values given in the table are based on $\log_{10}$ scale of flare intensity values.

Table A1
Mixed LSTM Regression Results of All Flares and M/X, B and C Flares Measured in MSE

	Num of hours before Event-Num of hours of data used							
Class	12-06	12-12	12-24	24-12	24-24	24-48	36-06	36-24
Average	0.25	0.25	0.24	0.25	0.27	0.28	0.29	0.30
M/X	0.44	0.46	0.48	0.61	0.63	0.69	0.72	0.71
C	0.19	0.20	0.19	0.14	0.19	0.16	0.15	0.15
B	0.25	0.23	0.22	0.29	0.25	0.27	0.26	0.28

Appendix B: Tables of Classification Results

The following six tables provide the summary results of all classification models, including M/X versus B (Tables B1 and B4), M/X versus B/Q (Tables B2 and B5), and M/X versus C/B/Q (Tables B3 and B6). Specifically, Tables B4–B6 provide the raw confusion matrices we obtained from experiments, while the summaries of different metrics in Tables B1–B3 are calculated based on them.

Table B1
M/X Versus B Flare Classification Results (Calculated Based On Table B4)

	Num of hours before Event-Num of hours of data used							
Metrics	12–06	12–12	12–24	24–12	24–24	24–48	36–06	36–24
Recall	0.89	0.89	0.91	0.80	0.80	0.80	0.74	0.74
Precision	0.92	0.92	0.93	0.89	0.92	0.91	0.94	0.94
F_1 score	0.91	0.91	0.92	0.85	0.85	0.85	0.82	0.82
HSS_1	0.82	0.81	0.84	0.71	0.72	0.72	0.68	0.69
HSS_2	0.86	0.86	0.88	0.75	0.78	0.76	0.71	0.71
TSS	0.85	0.85	0.88	0.74	0.76	0.75	0.69	0.70

Table B2
M/X Versus B/Q Flare Classification Results (Calculated Based On Table B5)

	Num of hours before Event-Num of hours of data used							
Metrics	12-06	12-12	12-24	24-12	24-24	24-48	36-06	36-24
Recall	0.91	0.89	0.90	0.79	0.80	0.80	0.74	0.74
Precision	0.64	0.66	0.66	0.72	0.71	0.68	0.68	0.66
F_1 score	0.75	0.75	0.76	0.75	0.75	0.73	0.70	0.69
HSS_1	0.39	0.42	0.43	0.48	0.46	0.39	0.34	0.31
HSS_2	0.73	0.74	0.74	0.73	0.72	0.70	0.67	0.66
TSS	0.88	0.86	0.87	0.76	0.77	0.76	0.70	0.70

Table B3
M/X Versus C/B/Q Flare Classification Results (Calculated Based On Table B6)

	Num of hours before Event-Num of hours of data used							
Metrics	12-06	12-12	12-24	24-12	4-24	24-48	36-06	36-24
Recall	0.54	0.49	0.45	0.35	0.34	0.32	0.29	0.32
Precision	0.45	0.47	0.47	0.54	0.52	0.53	0.55	0.56
F_1 Score	0.49	0.48	0.46	0.42	0.41	0.40	0.38	0.40
HSS_1	−0.11	−0.06	−0.05	0.05	0.02	0.03	0.06	0.07
HSS_2	0.47	0.45	0.44	0.39	0.38	0.37	0.35	0.37
TSS	0.51	0.46	0.43	0.33	0.32	0.30	0.28	0.30

Table B4
M/X Versus B Confusion Matrices

	Confusion matrix (mean [min, max])			
Model	TP	FN	FP	TN
12-06	86.2 [83,88]	8.8 [7,12]	7.3 [1,14]	176.7 [170,183]
12-12	84.2 [80,88]	10.8 [7,15]	6.8 [3,10]	177.2 [174,181]
12-24	85.4 [79,88]	9.6 [7,16]	6.4 [4,8]	177.6 [176,180]
18-06	79.5 [74,86]	10.5 [4,16]	7.9 [3,19]	156.1 [145,161]
18-12	79.2 [76,84]	10.8 [6,14]	5.4 [1,12]	158.6 [152,163]
18-24	81.1 [75,88]	8.9 [2,15]	7.9 [1,35]	156.1 [129,163]
24-06	71.7 [66,78]	17.3 [11,23]	4.3 [2,7]	158.7 [156,161]
24-12	70.3 [63,76]	18.7 [13,26]	5.2 [1,9]	157.8 [154,162]
24-24	71.0 [66,76]	18.0 [12,23]	6.8 [3,12]	156.2 [151,160]
24-48	64.4 [60,71]	16.6 [10,21]	6.4 [3,12]	113.6 [108,117]
36-06	57.5 [49,63]	20.5 [15,29]	4.1 [2,9]	89.9 [85,92]
36-12	59.9 [53,67]	18.1 [11,25]	6.8 [2,17]	87.2 [77,92]
36-24	57.6 [53,63]	20.4 [15,25]	4.1 [2,15]	89.9 [79,92]
36-48	59.4 [49,65]	18.6 [13,29]	6.1 [2,14]	87.9 [80,92]

Table B5
M/X Versus B/Q Confusion Matrices

	Confusion matrix (mean [min, max])			
Model	TP	FN	FP	TN
12-06	86.2 [83,88]	8.8 [7,12]	49.0 [29,73]	1,606.0 [1582,1626]
12-12	84.2 [80,88]	10.8 [7,15]	44.3 [33,55]	1,610.7 [1600,1622]
12-24	85.4 [79,88]	9.6 [7,16]	44.6 [35,57]	1,610.4 [1598,1620]
18-06	79.5 [74,86]	10.5 [4,16]	63.4 [23,113]	1,571.6 [1522,1612]
18-12	79.2 [76,84]	10.8 [6,14]	51.3 [27,78]	1,583.7 [1557,1608]
18-24	81.1 [75,88]	8.9 [2,15]	59.0 [21,167]	1,576.0 [1468,1614]
24-06	71.7 [66,78]	17.3 [11,23]	25.6 [18,33]	915.4 [908,923]
24-12	70.3 [63,76]	18.7 [13,26]	27.8 [14,40]	913.2 [901,927]
24-24	71.0 [66,76]	18.0 [12,23]	29.8 [20,40]	911.2 [901,921]
24-48	64.4 [60,71]	16.6 [10,21]	32.7 [17,57]	865.3 [841,881]
36-06	57.5 [49,63]	20.5 [15,29]	31.1 [8,80]	840.9 [792,864]
36-12	59.9 [53,67]	18.1 [11,25]	43.0 [19,99]	829.0 [773,853]
36-24	57.6 [53,63]	20.4 [15,25]	33.8 [13,100]	838.2 [772,859]
36-48	59.4 [49,65]	18.6 [13,29]	39.8 [21,78]	832.2 [794,851]

Table B6
M/X Versus C/B/Q Confusion Matrices

	Confusion matrix (mean [min, max])			
Model	TP	FN	FP	TN
12-06	49.8 [40,57]	45.2 [38,55]	54.7 [37,67]	1,998.3 [1986,2016]
12-12	47.1 [38,58]	47.9 [37,57]	53.5 [42,79]	1,999.5 [1974,2011]
12-24	41.6 [32,54]	53.4 [41,63]	44.7 [31,64]	2,008.3 [1989,2022]
18-06	36.6 [24,51]	53.4 [39,66]	35.5 [24,54]	1,856.3 [1838,1868]
18-12	37.3 [29,43]	52.7 [47,61]	31.7 [18,42]	1,860.3 [1850,1874]
18-24	35.0 [26,46]	55.0 [44,64]	29.2 [16,41]	1,862.8 [1851,1876]
24-06	32.2 [27,40]	48.8 [41,54]	30.7 [20,38]	1,137.3 [1130,1148]
24-12	29.4 [24,35]	51.6 [46,57]	26.1 [17,33]	1,141.9 [1135,1151]
24-24	28.8 [19,39]	52.2 [42,62]	27.7 [20,33]	1,140.3 [1135,1148]
24-48	28.0 [22,38]	53 [43,59]	25.1 [12,32]	1,142.9 [1136,1156]
36-06	23.6 [12,33]	54.4 [45,66]	17.0 [10,22]	1,025.0 [1020,1032]
36-12	26.9 [13,36]	51.1 [42,65]	19.9 [7,33]	1,022.1 [1009,1035]
36-24	25.2 [19,29]	52.8 [49,59]	21.5 [14,40]	1,020.5 [1002,1028]
36-48	25.1 [9,35]	52.9 [43,69]	15.9 [9,28]	1,026.1 [1014,1033]

Appendix C: Summary of Accuracy of Quiet Sample Prediction

Table C1 summarizes the classification results of quiet samples measured in two metrics. The definition of these two metrics are illustrated in section 3.3.

Table C1
Classification Results of Quiet Samples Measured in Accuracy

	Accuracy (mean [min, max] in %)			Accuracy	
Model	Metric 1	Metric 2	Model	Metric 1	Metric 2
12-06	99.4 [98.9,99.8]	89.0 [83.5,92.3]	24-12	99.6 [99.4,99.9]	88.2 [85.5,91.3]
12-12	99.4 [98.9,99.8]	89.1 [86.0,91.6]	24-24	99.5 [99.4,99.9]	87.6 [82.9,91.9]
12-24	99.6 [99.2,99.9]	88.9 [82.9,92.4]	24-48	99.5 [99.2,99.7]	86.8 [83.4,92.5]
18-06	99.1 [98.7,99.5]	87.4 [83.8,91.7]	36-06	99.6 [99.1,100]	88.6 [84.8,91.9]
18-12	99.3 [99.0,99.7]	88.6 [86.9,90.1]	36-12	99.4 [98.5,99.9]	87.5 [82.9,92.5]
18-24	99.4 [99.0,99.7]	88.8 [85.7,92.7]	36-24	99.3 [98.2,99.7]	84.1 [74.2,90.9]
24-06	99.3 [99.1,99.9]	86.6 [77.8,90.7]	36-48	99.5 [99.0,99.9]	87.6 [84.1,89.6]

Data Availability Statement

All SHARP data used in this study are available from the Joint Science Operations Center (JSOC) NASA grant; see https://jsoc.stanford.edu/. All relevant digital values used in the manuscript (both data and model) will be permanently archived at the U-M Library Deep Blue data repository, which is specifically designed for U-M researchers to share their research data and to ensure its long-term viability. To cite these data, please use the following format: Jiao, Z., Chen, Y., Manchester, W. (2020). Data for solar flare intensity prediction with machine learning models [data set]. University of Michigan—Deep Blue. https://doi.org/10.7302/b07j-bj08.

Acknowledgments

We thank Professors Gabor Toth, Tuija Pulkkinen, Shasha Zou, and Igor Sokolov from the Climate and Space Sciences and Engineering (CLaSP) at the University of Michigan for helpful comments and discussions. This work is supported by NSF Grant AGS-1322543 and NASA Grants 80NSSC19K0373 and 80NSSC18K1208. We also acknowledge support from the NASA DRIVE Center at the University of Michigan under NASA grant 80NSSC20K0600.

References

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. U.S. National Library of Medicine.

Baker, D. N., Balstad, R., Bodeau, J. M., Cameron, E., Fennell, J. F., Fisher, G. M., et al. (2009). Severe space weather events—Understanding societal and economic impacts workshop report. Committee on the Societal and Economic Impacts of Severe Space Weather Events, National Research Council, Washington, D.C.

Barnes, G., Leka, K. D., Schrijver, C. J., Colak, T., Qahwaji, R., Ashamari, O. W., et al. (2016). A comparison of flare forecasting methods. I. Results from the "all-clear" workshop. *The Astrophysical Journal*, *829*(2), 89. https://doi.org/10.3847/0004-637x/829/2/89

Bloomfield, D. S., Higgins, P. A., McAteer, R. T. J., & Gallagher, P. T. (2012). Toward reliable benchmarking of solar flare fore casting methods. *The Astrophysical Journal*, *747*(2), L41. https://doi.org/10.1088/2041-8205/747/2/l41

Bobra, M. G., & Couvidat, S. (2015). Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, *798*(2), 135. https://doi.org/10.1088/0004-637X/798/2/135

Bobra, M. G., Sun, X., Hoeksema, J. T., Turmon, M., Liu, Y., Hayashi, K., et al. (2014). The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs—Space-Weather HMI Active Region Patches. *Solar Physical*, *289*(9), 3549–3578.

Camporeale, E. (2019). The challenge of machine learning in space weather nowcasting and forecasting. *Space Weather*, *17*, 1179–1195. https://doi.org/10.1029/2018SW002061

Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., DuFumier, B., Zhou, T., et al. (2019). Identifying solar flare precursors using time series of SDO/HMI images and sharp parameters. *Space Weather*, *17*, 1404–1426. https://doi.org/10.1029/2019SW002214

Doan, C. D., & Liong, S.-Y. (2004). Generalization for multilayer neural network Bayesian regularization or early stopping. In *Proceedings of Asia Pacific Association of Hydrology and Water Resources 2nd Conference*, pp. 5–8.

Florios, K., Kontogiannis, I., Park, S.-H., Guerra, J. A., Benvenuto, F., Bloomfield, D. S., & Georgoulis, M. K. (2018). Forecasting solar flares using magnetogram-based predictors and machine learning. *Solar Physics*, *293*(2), 11–28. https://doi.org/10.1007/s11207-018-1250-4

Forbes, T. G. (2000). A review on the genesis of coronal mass ejections. *Journal Geophysical Research*, *105*(A10), 23,153–23,165.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press.

Gu, M., Mengi, E., Overton, M. L., Xia, J., & Zhu, J. (2006). Fast methods for estimating the distance to uncontrollability. *SIAM Journal on Matrix Analysis and Applications*, *28*(2), 477–502. https://doi.org/10.1137/05063060X

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, *Springer Series in Statistics*. New York, NY: Springer. Retrieved from https://books.google.com/books?id = eBSgoAEACAAJ

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Janocha, K., & Czarnecki, W. (2017). On loss functions for deep neural networks in classification. Schedae Informaticae 25.

Leka, K. D., & Barnes, G. (2018). Solar flare forecasting: Present methods and challenges.In N. Buzulukova (Ed.), *Extreme events in geospace* (pp. 65–98). Amsterdam, Netherlands: Elsevier. https://doi.org/10.1016/B978-0-12-812700-1.00003-0

Leka, K. D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., et al. (2019a). A comparison of flare forecasting methods. II. Benchmarks, metrics, and performance results for operational solar flare forecasting systems. *The Astrophysical Journal Supplement Series*, *243*(2), 36. https://doi.org/10.3847/1538-4365/ab2e12

Leka, K. D., Park, S.-H., Kusano, K., Andries, J., Barnes, G., Bingham, S., et al. (2019b). A comparison of flare forecasting methods. III. Systematic behaviors of operational solar flare forecasting systems. *The Astrophysical Journal*, *881*(2), 101. https://doi.org/10.3847/1538-4357/ab2e11

Liu, H., Liu, C., Wang, J. T. L., & Wang, H. (2019). Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, *877*(2), 121. https://doi.org/10.3847/1538-4357/ab1b3c

Maynard, T., Smith, N., & Gonzales, S. (2013). Solar storm risk to the North American electric grid: Lloyd's Insurance Company. https://www.lloyds.com/news-and-insight/risk-insight/library/natural-environment/solar-storm

Schrijver, C. J. (2009). Driving major solar flares and eruptions: A review. *Advance Space Research*, *43*(5), 739–755. https://doi.org/10.1016/j.asr.2008.11.004

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014a). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014b). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Space Weather Prediction Center (2019). Solar cycle progression. https://www.swpc.noaa.gov/products/solar%2Dcycle%2Dprogression, Accessed: 2019-11-16.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 3320–3328). Red Hook, NY: Curran Associates, Inc.