Draft version April 1, 2022
Typeset using LaTeX default style in AASTeX631

Predicting Solar Flares Using CNN and LSTM on Two Solar Cycles of Active Region Data

Zeyu Sun [1], Monica G. Bobra [2], Xiantong Wang [3], Yu Wang [4], Hu Sun [4], Tamas Gombosi [3], Yang Chen [4] and Alfred Hero [1,4]

[1] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105, USA
[2] W.W. Hansen Experimental Physics Laboratory, Stanford University, Stanford, CA 94305, USA
[3] Department of Climate and Space Sciences and Engineering, University of Michigan, Ann Arbor, MI 48109, USA
[4] Department of Statistics, University of Michigan Ann Arbor, MI 48109, USA

ABSTRACT

We consider the flare prediction problem that distinguishes flare-imminent active regions that produce an M- or X-class flare in the future 24 hours, from quiet active regions that do not produce any flare within ±24 hours. Using line-of-sight magnetograms and parameters of active regions in two data products covering Solar Cycle 23 and 24, we train and evaluate two deep learning algorithms—CNN and LSTM—and their stacking ensembles. The decisions of CNN are explained using visual attribution methods. We have the following three main findings. (1) LSTM trained on data from two solar cycles achieves significantly higher True Skill Scores (TSS) than that trained on data from a single solar cycle with a confidence level of at least 0.95. (2) On data from Solar Cycle 23, a stacking ensemble that combines predictions from LSTM and CNN using the TSS criterion achieves significantly higher TSS than the "select-best" strategy with a confidence level of at least 0.95. (3) A visual attribution method called Integrated Gradients is able to attribute the CNN's predictions of flares to the emerging magnetic flux in the active region. It also reveals a limitation of CNN as a flare prediction method using line-of-sight magnetograms: it treats the polarity artifact of line-of-sight magnetograms as positive evidence of flares.

1. INTRODUCTION

Solar flares are abrupt electromagnetic explosions occurring in magnetically active regions on the solar surface. Intense solar flares are frequently followed by coronal mass ejections and solar energetic particles, which may disturb or disable satellites, terrestrial communication systems, and power grids. Predicting such strong flares from solar observations is therefore of particular significance and has been one of the primary tasks in space weather research.

Flare prediction can be posed as a classification problem, asking for a binary decision on whether the sun will produce a flare above some level in a future time window. Since strong solar flares mostly occur in active regions, it is common to first produce predictions for each active region on the solar disk. In this paper, we consider a "strong-vs-quiet" flare prediction problem, distinguishing active regions that will produce an M- or X- class flare in the future 24 hours, from those that stay flare quiescent within 24 hours before—and after—the forecast issuance time.

Over the past decade, a great amount of flare prediction studies have been conducted on a data product named Space-Weather HMI Active Region Patches (SHARPs, Bobra et al. 2014). The SHARP database is derived from full-disk observations of the Helioseismic and Magnetic Imager (HMI, Schou et al. 2012) aboard the *Solar Dynamics Observatory* (SDO), containing maps and summary parameters of automatically tracked active regions from May 2010 to the present day, covering much of Solar Cycle 24. Despite the fact that SHARP is one of the most recent and highest quality datasets of its kind, it only contains a limited number of strong events, as Solar Cycle 24 is the weakest solar cycle in a century. Recently, a new data product, Space-Weather MDI Active Region Patches (SMARPs, Bobra et al. 2021), was developed as an effort to extend backward the SHARP database to include active region observations in Solar Cycle 23, a much stronger solar cycle with significantly more flaring events. In fact, Solar Cycle 23 is the longest

Corresponding author: Zeyu Sun
zeyusun@umich.edu

solar cycle (147 months) in the past 150 years[1]. The SMARP database was derived from the Michelson Doppler Imager (MDI, Scherrer et al. 1995) aboard the *Solar and Heliospheric Observatory* (SoHO), which observed the sun from 1996 to 2010. Compared to its successor HMI, MDI's measurement of the solar surface magnetic field is only restricted to the line-of-sight component, with lower spatial resolution, lower signal-to-noise ratio, and shorter cadence. As such, SMARP does not contain as much information as SHARP, and its data quality is not as high. Nonetheless, SMARP's coverage of a stronger solar cycle and its partial compatibility with SHARP make it a valuable data product to use with SHARP, especially for statistical studies in which a large sample size or a long time span is desired.

Many machine learning methods for flare prediction have been proposed in recent years. They roughly fall into three categories in terms of how flare pertinent features are extracted from data. The first category uses *explicit* parameterization of observational data that are considered relevant to flare production, e.g., SHARP parameters that characterize the photospheric magnetic field. Much of the effort in data-driven flare forecasting has been made in this category, exploring a wide range of machine learning algorithms including discriminant analysis (Leka & Barnes 2003), regularized linear regression (Jonas et al. 2018), support vector machine (Yuan et al. 2010; Bobra & Couvidat 2015; Nishizuka et al. 2017; Florios et al. 2018), k-nearest neighbors (Nishizuka et al. 2017), extremely random trees (Nishizuka et al. 2017), random forests (Liu et al. 2017; Florios et al. 2018), multi-layer perceptrons (MLP) (Florios et al. 2018), residual networks (Nishizuka et al. 2018, 2020), long short-term memory (LSTM) networks (Chen et al. 2019; Liu et al. 2019), etc. The second category learns features from images using fixed transformations, e.g., random filters (Jonas et al. 2018), Gabor filters (Jonas et al. 2018), wavelet transforms (Hada-Muranushi et al. 2016). The third category, only popularized more recently, *implicitly* learns flare indicative signatures directly from active region magnetic field maps. This category features mainly convolutional neural networks (CNNs) (Huang et al. 2018; Li et al. 2020). Note that the three categories are not mutually exclusive. For example, methods in the second category typically also depend on explicitly constructed features (e.g. Jonas et al. 2018) as the information within transformation coefficients is often limited. In this study, two representative deep learning methods, LSTM and CNN, are considered. LSTM uses times series of active region summary parameters derived from line-of-sight magnetograms, whereas CNN uses static point-in-time magnetograms.

With so many machine learning algorithms developed for flare forecasting, one might expect an improved performance by combining different methods. This expectation seems even more reasonable if component methods in the combination use different data to provide complementary information. This is the idea behind ensemble learning, a learning paradigm that capitalizes on different models to achieve better performance than is achievable by any of the models alone. During the past few decades, the rapidly-evolving field of ensemble learning has achieved great success in many areas, which has attracted the attention of the space weather community (Murray 2018). In this paper, using CNN and LSTM models independently trained on active region magnetograms and parameter sequences, respectively, we consider a particular type of ensemble method called stacking (Wolpert 1992). Ideas similar to our stacking ensemble method have previously appeared in solar flare forecasting, most notably Guerra et al. (2015, 2020). In their work, full-disk probabilistic forecasts are linearly combined with weights selected by maximizing a potentially non-convex performance metric (e.g., the True Skill Score, the Heidke Skill Score). In contrast, our stacking ensemble linearly combines two state-of-the-art machine learning classifiers (LSTM and CNN). To select the combination weights, we consider a convex cross-entropy loss function in addition to other performance metrics.

Deep learning models are often considered to be "black-box" due to lack of interpretability. Recently the machine learning community has developed empirical tools that aim to better interpret the decisions of the deep neural network (DNN). Among these tools is the class of attribution methods (e.g. Springenberg et al. 2015; Selvaraju et al. 2017; Shrikumar et al. 2017; Sundararajan et al. 2017) that attribute a score to gauge the contribution of each input feature for a given input sample. Attribution methods such as the occlusion method and Grad-CAM have been previously used to interpret CNNs in flare prediction applications (Bhattacharjee et al. 2020; Yi et al. 2021). In this work, we evaluate additional attribution methods (Deconvolution, Guided Backpropagation, DeepLIFT, and Integrated Gradients) on the interpretation of CNNs trained to predict flares. In particular, we show that Integrated Gradient attribution maps, which have the same resolution as the input image, lead to insights on the important magnetic features that inform the CNN's decisions on flare prediction.

The contributions of this paper are as follows:

[1] Source: https://ntrs.nasa.gov/api/citations/20130013068/downloads/20130013068.pdf

1. We demonstrate the value of combining SMARP and SHARP to improve flare prediction performance.

2. We compare the flare prediction performance of LSTM and CNN on an equal footing, i.e., on the same temporally evolving active region dataset.

3. We demonstrate that stacking the LSTM and CNN can significantly improve flare class prediction in certain settings.

4. We provide visual explanations of the CNN predictor using visual attribution methods including Deconvolution, Guided Backpropagation, Integrated Gradients, DeepLIFT, and Grad-CAM. We demonstrate the potential of these methods in identifying flare indicative signatures, interpreting CNN's decisions, revealing model limitations, and suggesting model modifications.

The rest of the paper is organized as follows. Section 2 introduces the data sources and how they are processed into machine-learning-ready datasets. Section 3 describes the flare prediction methods, stacking ensemble, and visual attribution methods. Section 4 presents and compares the flare prediction performance on the datasets. Section 5 concludes the paper by presenting the lessons learned from the experiments. Our codes for data processing, model training, and performance evaluating are openly available at https://github.com/ZeyuSun/flare-prediction-smarp.

2. DATA

2.1. *Data sources*

Observational data of active regions of Solar Cycle 23 and 24 are extracted from the SMARP and the SHARP data product, respectively. Both SMARP and SHARP contain automatically-tracked active region cutouts of full-disk line-of-sight magnetograms, referred to as Track Active Region Patches (TARPs) and HMI Active Region Patches (HARPs), respectively. They also contain summary parameters that characterize physical properties of active regions. We consider definitive SMARP and SHARP records in Cylindrical Equal-Area (CEA) coordinates hosted in Joint Science Operations Center[2]. We query SMARP records from 1996 April 23 to 2010 October 28 and SHARP records from 2010 May 1 to 2020 December 1, both at a cadence of 96 minutes. Only good quality SMARP and SHARP records within $\pm 70^\circ$ of the central meridian matching at least one NOAA active region are considered. For active region summary parameters, we use four metadata keywords that are common to SMARP and SHARP, i.e., `USFLUXL`, `MEANGBL`, `R_VALUE`, and `AREA`. Definitions of these summary parameters are listed in Table 1. For images, we use photospheric line-of-sight magnetic field maps, or magnetograms, from the two data products.

Observational data samples are labeled using the GOES catalog of X-ray solar flare events. Based on the peak magnitude of 1–8 Å soft X-ray flux measured by *Geostationary Operational Environmental Satellites* (GOES), solar flare events are classified into five increasingly intense classes: A, B, C, M, and X, sometimes appended with a number that indicates a finer scale classification. M- and X- class flares are referred to as strong flares throughout the paper.

Table 1. Active region summary parameters used in this study. The line-of-sight magnetic field is denoted by B_L.

Keyword	Description	Pixels	Formula	Unit
`USFLUXL`	Total line-of-sight unsigned flux	Pixels in the TARP/HARP region	$\sum \lvert B_L \rvert dA$	Maxwell
`MEANGBL`	Mean gradient of the line-of-sight field	Pixels in the TARP/HARP region	$\frac{1}{N}\sum\sqrt{\left(\frac{\partial B_L}{\partial x}\right)^2+\left(\frac{\partial B_L}{\partial y}\right)^2}$	Gauss/pixel
`R_VALUE`	R, or a measure of the unsigned flux near polarity inversion lines (Schrijver 2007)	Pixels near polarity inversion lines	$\log\left(\sum \lvert B_L \rvert dA\right)$	Maxwell
`AREA`	De-projected area of patch on sphere in micro-hemisphere	Pixels in the TARP/HARP region	$\sum dA$	mH

[2] See http://jsoc.stanford.edu.

We only consider GOES solar flare events with at least one associated NOAA active region that can be used to cross-reference the `NOAA_ARS` keyword in SHARP (or SMARP) databases to associate the flare with a HARP (or TARP). The GOES event records are queried using the Sunpy package (The SunPy Community et al. 2020) from the beginning of 1996 to the end of 2020, covering the period of the SMARP and SHARP observations used in this paper. Of note, although the GOES catalog is widely considered as the "go-to" record database in solar flare forecasting, it is not error-free. There are cases in which flares, even the strong ones, are not assigned to any active region (Leka et al. 2019). Furthermore, small-sized flares could be buried under the background radiation, a phenomenon frequently observed for A- and B-class flares, especially after a strong flare occurs. Moreover, there are 61 event records annotated with an unknown GOES event class, most of them in the year 1996. These 61 event records are excluded in this study.

2.2. *Data fusion*

The challenge of combining the disparate SHARP and SMARP data is mitigated by the fact that there is a short overlapping time period over which they were jointly collected (May 1 to October 28 of 2010). We used this common time period to evaluate the dissimilarities between the SHARP/SMARP data and to develop methods for data alignment. As explained below, our analysis of the data over the common time period led us to adopt a simple method for fusing the SHARP and SMARP data: (1) we downsampled the SHARP magnetograms to match the resolution of the SMARP magnetograms; (2) we separately transformed the SHARP and SMARP summary parameters by Z-score (translation-scale) standardization.

We first discuss the fusion of the SHARP and SMARP magnetograms. SHARP magnetograms inherit the HMI resolution of about $0.5''$ per pixel, whereas SMARP magnetograms inherit the MDI resolution of about $2''$ per pixel. To compare HMI and MDI magnetograms, Liu et al. (2012) reduced HMI spatial resolution to match MDI's by convolving a two-dimensional Gaussian function with an FWHM of 4.7 HMI pixels and truncated at 15 HMI pixels. Then, the HMI pixels enclosed in each MDI pixel are averaged to generate an MDI proxy pixel. Subsequently, a pixel value transformation $\mathrm{MDI} = -0.18 + 1.40 \times \mathrm{HMI}$ is applied. In this work, we adopted a simpler approach by subsampling SHARP magnetograms 4 times in both dimensions to match the resolution of SMARP magnetograms. Unlike Liu et al. (2012), we approximated pixel value transformation with an identity map, as pixel value distributions of SHARP and SMARP magnetograms in the overlap period are very similar (Figure 1). Our approximation agrees well with Riley et al. (2014), who found a multiplicative conversion factor of 0.99 ± 0.13 between MDI and HMI using histogram equating. The discrepancy between our multiplicative conversion factor (1.099) and that of Liu et al. (2012) (1.40) may be because they considered full-disk magnetograms whereas we focus on active regions. In addition, they considered only 12 pairs in June–August 2010, whereas we considered every possible matching in May–October 2010. Moreover, they performed pixel-to-pixel match of full-disk magnetograms, whereas we use histogram-based methods on active regions because more precise models for aligning coordinates between CEA-projected SHARP and SMARP are not yet available (Bobra et al. 2021). Furthermore, they considered pixels within 0.866 solar radius of Sun's center, whereas we considered pixels within $\pm 70^\circ$ from the central meridian.

We next discuss the fusion of SHARP and SMARP summary parameters. Although designed to represent the same physical quantity, summary parameters with identical keywords in SHARP and SMARP are calculated from two pipelines with different source data, and the differences between them cannot be neglected. Bobra et al. (2021) investigated these differences by comparing the marginal and the pairwise joint distribution of co-temporal SMARP and SHARP summary parameters for 51 NOAA active regions over the overlap period of MDI and HMI (Bobra et al. 2021, Figure 3). Motivated by these findings, we investigated the linear associations between SHARP and SMARP using a univariate linear regression analysis. Specifically, SMARP parameters were regressed on their SHARP counterparts. As shown in Figure 2, `USFLUXL` is the most correlated parameter between SHARP and SMARP, with Pearson correlation coefficient $r = 0.970$, whereas `MEANGBL` is the least correlated parameter, with $r = 0.796$. Note that applying linear transformations to the SHARP summary parameters would have no effect once Z-score standardization was performed. This is because Z-scores are invariant to univariate linear transformation. Therefore, in practice, the linear transformation on SHARP summary parameters is not performed.

2.3. *Sample extraction and labeling*

We focus our joint SMARP and SHARP analysis on a particular task of interest, which we refer to as "strong-vs-quiet" flare prediction: based on a sequence of observations of an evolving active region, the objective is to discriminate active regions that will generate strong flares in the near future, from active regions having no flare activity whatsoever.

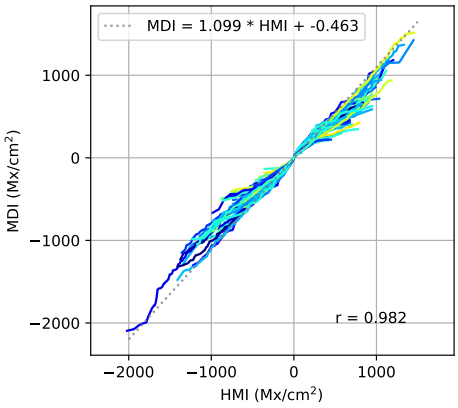


Figure 1. Q-Q (quantile-quantile) plot of 50 matched magnetogram pairs of HARP and TARP from May 1 to October 28 in 2010. Active regions with pixels outside of $\pm 70^\circ$ from the central meridian are not used. For each pair, the co-temporal magnetograms are sampled at a rate of every 8 hours. The pixels within the intersection of the bounding boxes of active region pairs are used. Lighter colors indicate higher latitudes.

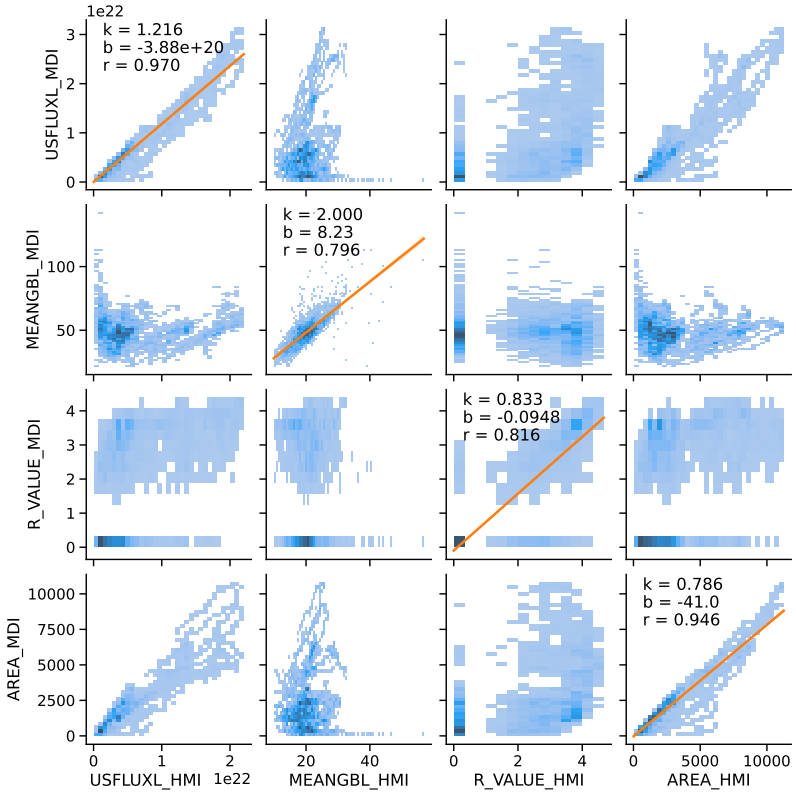


Figure 2. 2D histograms of summary parameters `USFLUXL`, `MEANGBL`, `R_VALUE`, and `AREA` between SHARP and SMARP. SHARP summary parameters are suffixed with `_HMI` and SMARP with `_MDI`. The orange lines in the diagonal blocks are the least square fit of SMARP summary parameters on the corresponding SHARP summary parameters, with coefficient k, intercept b, and Pearson correlation coefficient r displayed in the corner.

Table 2. *Left*: Samples sizes of all possible flare activity evolution types, with missing/inconsistent data removed. The flare activity in the observation period of a sample can be `QUIET` (the active region is flare-quiet), `WEAK` (only flares of size smaller than M1.0 occur), or `STRONG` (there is at least one large flare of size M1.0 or above). The prediction period can be classified the same way. The entries denote the sample counts in SMARP/SHARP data sets. *Right*: The associations of flare activity evolution types and the class labels for the "strong-vs-quiet" flare prediction task. Positive samples are denoted as `+` and negative samples as `-`. Samples with the evolution type signifying a decaying flare activity (denoted as a blank space) or leading to only small flares (denoted as `?`) are not relevant to our task.

	Prediction		
Observation	QUIET	WEAK	STRONG
QUIET	130695 / 66349	12341 / 10715	932 / 296
WEAK	12688 / 11110	12033 / 14891	1915 / 1366
STRONG	1071 / 282	1723 / 1371	1754 / 1187

	Prediction		
Observation	QUIET	WEAK	STRONG
QUIET	-	?	+
WEAK		?	+
STRONG			+

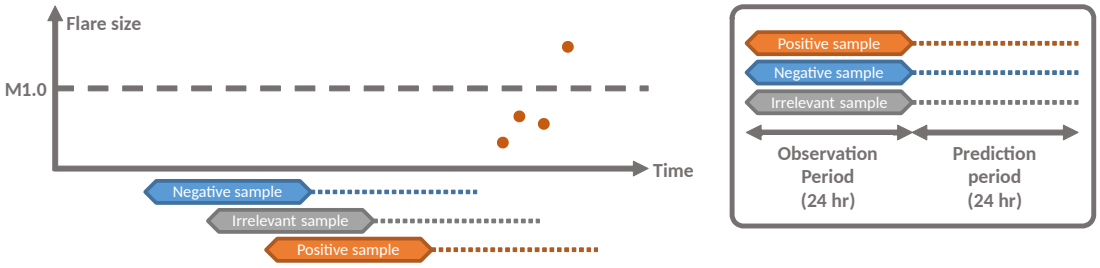


Figure 3. Demonstration of the sample extraction and labeling procedure of an active region. The dark orange dots represent flares that occurred in an active region, with the last flare exceeding the M1.0 threshold. The blue sample is labeled as the negative class because no flare of any class occurs in the observation and the prediction period. The gray sample is irrelevant to the task since all flares in the prediction period are weaker than M1.0. The orange sample is labeled as the positive class because the prediction period contains a flare of size exceeding M1.0.

To construct a dataset for this task, we extract data samples using a sliding window approach similar to Angryk et al. (2020). Specifically, samples are extracted from a 24-hour time window that slides through an active region sequence with step size of 96 minutes (i.e., one 24-hour subsequence starts 96 minutes after the starting point of its previous subsequence). The 24-hour time window that a sample covers is called the *observation period*, and the 24-hour time window following immediately after the observation period is called the *prediction period*. Then we retain the samples that either: (1) exhibit an M- or X-class flare in the prediction period (assigned to the positive class); or (2) have no flare of any class in both observation and the prediction period (assigned to the negative class). Table 2 lists the sample sizes of all the possible flare activity evolution types and the associations between evolution types and class labels. Figure 3 shows examples of extracted and labeled samples.

We note that two evolution types are excluded in this study. Evolution types denoted by blank spaces in Table 2 indicate a decay in flare activity—a process different from the onset or the continuation of the flare activity. They are unrelated to our task and also less studied in the literature. Including them in the dataset brings an unnecessary source of heterogeneity and makes learning a reliable predictor substantially more difficult. Evolution types denoted by question marks have only weak flares in the prediction period. They are excluded to enhance the contrast between the two classes, to avoid the concerns on the detection of weak flares (many B- and C-class flares are obscured by background radiation), and to avoid the controversy on the granularity of labels (for instance, an M1.0 class flare and a C9.9 class flare relieve a similar amount of energy but are categorized differently). Possible limitations of our sample selection rules are discussed in Section 5.

After extracting and labeling active region samples, we discarded the samples having inconsistent or missing data. We consider a point-in-time record in a sample sequence as a "bad record" if (1) the magnetogram contains Not-a-Number (NaN) pixels, (2) the magnetogram has either height or width deviating more than 2 pixels from the median dimension in the sample sequence, or (3) any of the summary parameters is NaN. A sample is discarded if it contains

Table 3. Sample sequences extracted from SMARP and SHARP

	Positive (M1.0+)	Negative (Quiet)	Event Rate
SMARP	4601	130695	0.0340
SHARP	2849	66349	0.0412

more than 2 bad records or the last record is a bad record. The validity of the last record is enforced because the CNN uses only the last record in the sample sequence.

The numbers of SMARP and SHARP samples output by the above pipeline are shown in Table 3. The count of negative samples is observed to dominate in both SMARP and SHARP. To address the issue of significant class imbalance, we randomly undersample the negative samples to equalize the positive and negative classes, which will be described in more detail in Section 2.5.

2.4. *Train/validation/test split*

A common practice in machine learning is to divide the data samples into three disjoint subsets, as known as splits: a training set on which the model is fitted, a validation set on which hyperparameters are selected, and a test set on which the model is evaluated for generalization performance. The ability of a trained machine learning algorithm to generalize to unseen samples hinges on the distributional similarity among splits. Therefore, it is important that splits be sufficiently similar in distribution.

Due to the temporal coherence of an active region in its lifetime, a random split of data samples will have samples coming from one active region categorized into different splits. Such correlation constitutes an undesirable information leakage among splits. For instance, information leaking from the training set into the test set will likely result in an overly optimistic estimate of the generalization performance. Much of the flare prediction literature deals with this issue by taking a chronological split, e.g., a year-based split (e.g. Bobra & Couvidat 2015; Chen et al. 2019). Unfortunately, it is observed that the splits may not share the same distribution due to solar cycle dependency (Wang et al. 2020). Some other works take an active-region-based split, where data samples from the same active region must belong to the same split (e.g. Guerra et al. 2015; Campi et al. 2019; Zheng et al. 2019; Li et al. 2020). Compared to splitting by years, this approach has the advantage that active regions in each split are randomly dispersed in different phases of a solar cycle, removing the bias introduced by artificially specifying splits. This distributional consistency between splits comes at the price of an additional source of information leakage due to sympathetic flaring in co-temporal active regions.

2.5. *Random undersampling*

As shown in Table 3, both SMARP and SHARP exhibit prominent class imbalance, with positive (minority class) samples significantly outnumbered by negative (majority class) samples. In flare forecasting, class imbalance has been recognized as a major challenge, both in forecast verification (Woodcock 1976; Jolliffe & Stephenson 2012) and in data-driven methodology (Bobra & Couvidat 2015; Ahmadzadeh et al. 2021). A data-driven forecasting method needs to be calibrated to handle class imbalance properly, in order to effectively detect the events of interest, as opposed to being overwhelmed by the sheer volume of the negative samples in the training set.

Methods to tackle class imbalance can be categorized into three types: data-level methods, algorithm-level methods, and a hybrid of the two (Krawczyk 2016; Johnson & Khoshgoftaar 2019). Data-level methods rebalance the class distribution by oversampling the minority class and/or undersampling the majority class—both have been used in flare forecasting (e.g. Ribeiro & Gradvohl 2021; Yu et al. 2010). Classifiers trained on rebalanced datasets, without being biased towards the majority class, are more likely to effectively detect the event of interest. Such classifiers are also generally more robust to variations in class imbalance than classifiers trained on the original imbalanced data (Xue & Hall 2014). Algorithm-level methods modify the learners to alleviate their bias towards the majority groups. The most popular algorithm-level approach—also widely used in flare forecasting (e.g. Bobra & Couvidat 2015; Nishizuka et al. 2018; Liu et al. 2019)—is cost-sensitive learning, which assigns a higher penalty to the misclassification of samples from the minority class to boost their importance (Krawczyk 2016). The penalty weights for different classes are usually set to be inversely proportional to their samples sizes (Nishizuka et al. 2018; Liu et al. 2019; Ahmadzadeh et al. 2021). Other algorithm-level approaches include imbalanced learning algorithms, one-class learning, and ensemble methods (Ali et al. 2013), but they are rarely used in flare forecasting. Recent work by Ahmadzadeh et al. (2021) provided

a thorough investigation of class imbalance in flare forecasting by presenting an empirical evaluation of multiple approaches.

We handle the class imbalance problem using random undersampling: we randomly remove samples from the majority class until the number of positive and negative samples are equalized. By training on rebalanced training and validation sets, we obtain a predictor that is more robust to shifts in climatological flare rates and learns more resilient pre-flare features. Following Zheng et al. (2019) and Deng et al. (2021), we also perform random undersampling on test sets to preserve the class proportion consistency among splits. By testing on rebalanced test sets, we evaluate the generalization performance of the predictor under the *same* climatological rate as it is trained with. We note there are also studies that do not rebalance the test set in order to evaluate the performance under a realistic event rate (Cinto et al. 2020; Ahmadzadeh et al. 2021). However, the bias caused by the class-balance change between the test and the training set is often neglected. We discuss such bias in addition to possible corrective methods in Section 5. Applying such corrections will be left to future work that directly addresses operational applications.

2.6. *Image resizing*

The CNN requires all input images to be of the same size, but the active region cutouts are of different sizes and aspect ratios. Resizing (via interpolation), zero padding, and cropping are among mostly used methods to convert different-sized images into a uniform size. Jonas et al. (2018) cropped and padded input images to a square aspect ratio and then downsampled them to 256×256 pixels. This has the advantage of preserving the aspect ratio. However, since many active regions are east-west elongated, cropping may exclude part of active regions and padding may introduce artificial signals. In this work, we resize all active region magnetograms to 128×128 pixels using bilinear interpolation, similar to Huang et al. (2018) and Li et al. (2020).

2.7. *Standardization*

Magnetogram pixel values and summary parameters are different physical quantities expressed in different units and ranges. Unlike physical modeling, many machine learning algorithms are invariant to scaling the input; they only care about the relative feature amplitudes. Moreover, drastically different ranges of features may hurt the convergence and stability of many algorithms. Therefore, the data of different scales are typically transformed into the same range via a process called standardization. In particular, Z-score standardization transforms the input data by removing the mean and then dividing by the standard deviation.

In this work, we apply the Z-score standardization to the image data using the mean and standard deviation of the magnetogram pixels in SHARP. This is because the pixel values between SMARP and SHARP are similar. We apply the Z-score standardization to SMARP and SHARP summary parameters separately. That is, the mean and standard deviation are calculated for SHARP and SMARP separately, and data in one dataset is standardized using the mean and the standard deviation in that dataset. The transformation is "global" (Ahmadzadeh et al. 2021) in that it is calculated regardless of the splits. Empirical evaluation in Ahmadzadeh et al. (2021) showed a global standardization to be better than the local standardization, i.e., the mean and standard deviation are calculated only for the training split. We note that, with this standardization, the linear transformation converting SHARP summary parameters to SMARP proxy data is no longer needed; any coefficients and bias will have no effect after standardization.

3. METHODOLOGY

In this section, we introduce the two deep learning models, LSTM and CNN, that we use for flare prediction. Then we describe the stacking ensemble approach that combines the two models. Subsequently, we describe the forecast verification methods (skill scores and graphical tools) we used. Then we discuss how we use the paired t-test to compare empirical performance between algorithms and settings with statistical confidence. Lastly, we introduce the visual attribution methods used to interpret the decisions generated by the CNN.

3.1. *Deep learning models*

We use two deep neural network models, LSTM and CNN, to predict strong flares from active region observation. LSTMs use 24-hour-long time series of summary parameters before the prediction period begins, whereas CNNs use the static point-in-time magnetogram right before the prediction period begins. Both networks output the probability that an input sample belongs to the positive class, i.e., the probability that the active region will produce a strong flare in the next 24 hours, rather than continue to be flare-quiescent.

The long short-term memory (LSTM) network was introduced by Hochreiter & Schmidhuber (1997) as a type of recurrent neural networks that learns from sequential data for classifying text and speech. A common LSTM unit is composed of a cell, an input gate, an output gate, and a forget gate. In solar flare prediction, LSTMs have been applied to prediction using SHARP parameter series (Chen et al. 2019; Liu et al. 2019). The architecture of the LSTM used in this paper is adapted from Chen et al. (2019), shown in Figure 4(a). Two LSTM layers, each with 64 hidden states, are stacked. The last output of the second LSTM layer, a 64-dimensional vector, is sent to a linear layer with 2 outputs. The softmax is applied to this 2-dimensional output to get the predicted probabilities of the positive and the negative class.

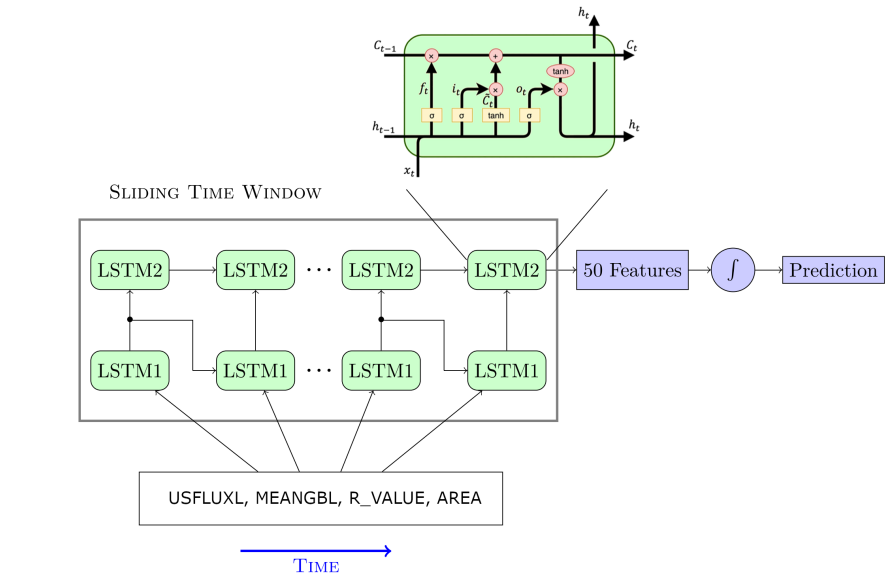
The convolutional neural network (CNN) is a neural network architecture that learns from images. CNNs have been applied to solar flare forecasting by Huang et al. (2018) and Li et al. (2020). We use the architecture proposed by Li et al. (2020), illustrated in Figure 4(b), which was itself inspired by the VGG network (Simonyan & Zisserman 2014) and the Alexnet network (Krizhevsky et al. 2012). The first two convolutional layers have kernels of size 11×11, designed to learn low-level and concrete features. The three following convolutional layers have kernels of size 3×3, designed to learn more high-level, abstract concepts. Batch normalization is used after all convolutional and linear layers to speed convergence. ReLU nonlinearity is applied to only convolutional layers. The batch normalization outputs of the two linear layers are randomly dropped out with a probability of 0.5 in training to reduce overfitting. The 2-dimensional output is passed to softmax to generate a probability assignment between the positive and the negative class. More details of this architecture can be found in Li et al. (2020).

The procedures used to train the LSTM and the CNN are similar. For both models, the Adam optimizer (Kingma & Ba 2014) is used to minimize the cross-entropy loss with learning rate 10^{-3} and batch size 64. Both models are evaluated on the validation set after each epoch of training. To prevent overfitting, the training is early-stopped if no improvement on the validation True Skill Score (or TSS, explained later in Section 3.3) is observed for a certain number of epochs, called the *patience*, before early stopping. The LSTM is trained for at most 20 epochs with a patience of 5 epochs, whereas the CNN is trained at most 20 epochs with a patience of 10 epochs. After training, the LSTM or the CNN with the best validation TSS among the checkpoints of all epochs is selected and evaluated on the test set to estimate its generalization performance.

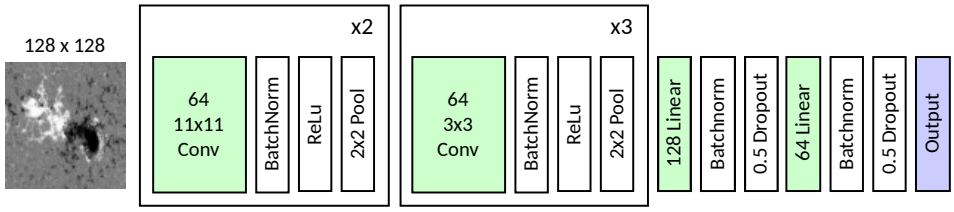
3.2. *Stacking ensemble*

In ensemble learning, the most common approach to combining individual models—called *base learners*—is averaging their outputs, possibly with non-equal weights, to produce a final output. Another approach is stacking. Different from output averaging, stacking uses cross-validation training to learn the best combination—called the *meta-learner*—of the outputs of the base learners. The meta-learner is often chosen to be global and smooth (Wolpert 1992), such as linear models (Breiman 1996; LeBlanc & Tibshirani 1996; Ting & Witten 1999) and decision trees (Todorovski & Džeroski 2003; Džeroski & Ženko 2004). Training a stacking ensemble consists of two stages. First, the base learners are fitted on the training set. Then, the predicted probabilities by all base learners on the validation set, as well as their labels, are collected into the so-called *level-one* data, on which the meta-learner is fitted. Cross-validation is frequently used in place of a simple train-validation split to significantly increase the sample size of the level-one dataset. Either way, it is important that the level-one data are out-of-sample for base learners, otherwise the meta-learner will inevitably prefer models that overfit the training data over ones that make more realistic decisions (Witten et al. 2016).

An early application of stacking for solar flare prediction problems was presented by Džeroski & Ženko (2004). These authors proposed a decision-tree-based stacking method, demonstrating it on the *UCI Repository of machine learning databases* (Dua & Graff 2017), including a dataset with 1389 flare instances, each characterized by 10 categorical attributes. In the space weather community, Guerra et al. (2015) proposed stacking for flare prediction. They linearly stacked four full-disk probabilistic forecasting methods, with the weights maximizing HSS under the constraint that they are non-negative and sum to 1. They found that the stacking ensemble performed similarly to an equally weighted model. Guerra et al. (2020) continued in this direction adopting stacking for more forecasting methods and a larger data sample. They also considered an unconstrained linear combination with a climatological frequency term. They found most ensembles perform better than a bagging model that essentially averages the members' predictions. However, these authors overlooked the nonconvexity of the objective in training the meta-learner. Furthermore, their conclusions about the superiority of the stacking ensemble over the equally weighted model were limited to the evaluation of in-



(a) LSTM architecture



(b) CNN architecture

Figure 4. Neural network architectures. (a) shows the LSTM architecture. (b) shows the CNN architecture.

sample error, which is unlikely to generalize. We will discuss these issues as we introduce our proposed stacking ensemble.

We formulate the stacking ensemble as a linear combination of LSTM and CNN. Given a sample x_i, the stacking ensemble outputs the probability that the sample belongs to the positive class

$$r_i = \alpha p_i + (1 - \alpha) q_i, \quad 0 \leq \alpha \leq 1, \tag{1}$$

where p_i and q_i are the predicted probability by LSTM and CNN, respectively, and α is the meta-learner parameter. We note that, in order to be most effective, stacking methods require base learners that are diverse and complement each other. Examples include base learners trained on different types of data, each providing an alternative view of the same phenomenon. The magnetograms and summary parameters in SHARP/SMARP provide such diverse multiviews. These multiviews are processed by CNN and LSTM, respectively, to generate two predicted probabilities, which are then fused into a single prediction by the aforementioned stacking procedure.

The meta-learning of the stacking ensemble essentially means finding the optimal combination weight α. To that end, we minimize a loss function that penalizes the difference between the prediction r_i and the binary label $y_i \in \{0, 1\}$ for samples in the validation set. A natural choice is to maximize the accuracy or skill scores, or equivalently, to minimize the loss which is the negation of these metrics. The downside of these loss functions is that they may not be convex or differentiable. This is not problematic when there are only a few base learners, as is the case with this work and Guerra et al. (2015); a grid search can be applied to find the weights that minimize the loss. However, as the number of base learners increases, the grid search quickly becomes infeasible, and iterative algorithms have to be used—many of them require convexity and differentiability of the loss function for guaranteed convergence. In machine learning, convexity and smoothness ensure the uniqueness of the minimizer and guarantee faster convergence rates for iterative algorithms (Nocedal & Wright 2006; Bottou et al. 2018). Guerra et al. (2020) found that for some optimization metrics, the

		Predicted		
		Negative	Positive	Total
True	Negative	TN	FP	N
	Positive	FN	TP	P
	Total	N′	P′	N + P

Table 4. A contingency table consisting of TP (true positive), FP (false positive), FN (true negative), and TN (true negative).

resulting weights were sensitive to the initialization of the solver. This is likely the consequence of the nonconvexity of the loss function. Their proposed solution was to run the solver with multiple initializations and take the average.

To circumvent convergence problems, machine learning researchers often use loss functions that are convex and differentiable. One example is the negative log-likelihood function for the logistic regression model, whose minimum corresponds to the maximum likelihood estimator (MLE). Within the meta-learning framework specified by Equation (1), the negative log-likelihood loss function is

$$L(\alpha) = -\log \prod_{i=1}^{n} r_i^{y_i} (1 - r_i)^{1-y_i} \tag{2}$$

$$= \sum_{i=1}^{n} \underbrace{\left(-y_i \log r_i - (1 - y_i) \log(1 - r_i)\right)}_{L_i} . \tag{3}$$

The negative log-likelihood objective can also be interpreted as the binary cross-entropy loss, a divergence measure between the distributions of ground truth labels and predicted probabilities. This loss function can be decomposed into the summation of instance-wise loss L_i, having gradient and the Hessian

$$L_i'(\alpha) = \left(-\frac{y_i}{r_i} + \frac{1 - y_i}{1 - r_i}\right)(p_i - q_i)\,, \tag{4}$$

$$L_i''(\alpha) = \left(\frac{y_i}{r_i^2} + \frac{1 - y_i}{(1 - r_i)^2}\right)(p_i - q_i)^2 \geq 0. \tag{5}$$

Minimizing L on $\alpha \in [0, 1]$ is a convex optimization problem and we use grid search to find the minimizer. In general cases with more than two base learners, iterative algorithms like projected gradient descent or Newton's method will be more efficient. We point out that convex loss functions are widely adopted in the literature of stacking, such as least square estimate (Breiman 1996), regularized linear regression (LeBlanc & Tibshirani 1996), multi-response linear regression (Ting & Witten 1999), and hinge loss (ŞEn & Erdogan 2013).

3.3. *Evaluation tools*

The prediction probabilities output by CNN and LSTM can be turned into binary decisions by thresholding and the algorithm performance can be represented as a contingency table (or confusion matrix), as shown in Table 4. The contingency table contains the most complete information for categorical prediction. However, a single numerical metric is often needed to summarize the table for model selection. Accuracy and skill scores are examples of such contingency table based metrics that are adopted in space weather forecasting.

We start our discussion on metrics with accuracy (ACC), also known as rate correct, the simplest metric that is widely used in all sorts of domains. In terms of the contingency table, accuracy is defined as

$$A = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{N} + \mathrm{P}}. \tag{6}$$

For a highly imbalanced classification problem like solar flare prediction, accuracy is generally not considered a useful metric, since a no-skill classifier that assigns the majority label to all samples will be correct most of the time. Therefore, a plethora of skill scores are devised to overcome this issue.

A skill score provides a normalized measure of the improvement against a specific reference method. In its most general form, a skill score can be expressed as

$$\text{Skill} = \frac{A_{\text{forecast}} - A_{\text{reference}}}{A_{\text{perfect}} - A_{\text{reference}}}, \tag{7}$$

where A_{forecast}, $A_{\text{reference}}$, and A_{perfect} are the accuracy of the forecast to be evaluated, the reference forecast, and the perfect forecast, respectively. A higher skill score indicates better performance, with the maximum value 1 corresponding to the perfect performance, 0 corresponding to no improvement over the reference, and negative values corresponding to performance worse than the reference. Below, we review some of the mostly used skills scores in flare forecasting. For a more complete discussion, we refer readers to Woodcock (1976) and Wilks (2011).

The Heidke Skill Score (HSS), also known as Cohen's kappa coefficient due to Cohen (1960), uses a random forecast independent from the flare occurrences as a reference. The expected number of correct forecasts made by the random predictor, denoted by E, can be calculated using the law of total expectation as

$$\mathrm{E} = \frac{\mathrm{P}}{\mathrm{N}+\mathrm{P}} \times \mathrm{P}' + \frac{\mathrm{N}}{\mathrm{N}+\mathrm{P}} \times \mathrm{N}'. \tag{8}$$

The accuracy of the random predictor can then be expressed as

$$A_{\text{reference}} = \frac{\mathrm{E}}{\mathrm{N}+\mathrm{P}}. \tag{9}$$

Defined using this reference accuracy, HSS has the form

$$\mathrm{HSS} = \frac{\mathrm{TP}+\mathrm{TN}-\mathrm{E}}{\mathrm{N}+\mathrm{P}-\mathrm{E}} = \frac{2[(\mathrm{TP}\times\mathrm{TN}) - (\mathrm{FN}\times\mathrm{FP})]}{\mathrm{PN}' + \mathrm{P}'\mathrm{N}}. \tag{10}$$

HSS quantifies the forecast improvements over a random prediction. Since the random reference forecast is dependent on the event rate (climatology) $\mathrm{P}/(\mathrm{N}+\mathrm{P})$, HSS has to be used with discretion in comparing methods when the event rate varies.

The True Skill Score (TSS), also known as Hanssen & Kuiper's Skill Score (H&KSS) or Peirce Skill Score. It is the difference between the probability of detection (POD) and the false alarm rate (FAR):

$$\mathrm{TSS} = \underbrace{\frac{\mathrm{TP}}{\mathrm{P}}}_{\mathrm{POD}} - \underbrace{\frac{\mathrm{FP}}{\mathrm{N}}}_{\mathrm{FAR}}. \tag{11}$$

TSS falls into the general skill score definition with a reference accuracy (Barnes et al. 2016)

$$A_{\text{reference}} = \frac{\mathrm{FN}(\mathrm{TN}-\mathrm{FP})^2 + \mathrm{FP}(\mathrm{TP}+\mathrm{FN})^2}{(\mathrm{N}+\mathrm{P})[\mathrm{FN}(\mathrm{TN}-\mathrm{FP}) + \mathrm{FP}(\mathrm{TP}+\mathrm{FN})]}, \tag{12}$$

constructed such that both the random and unskilled predictors score 0. A nice property of TSS is its invariance to the class imbalance ratio, and hence is suggested by Bloomfield et al. (2012) to be the standard measure for comparing flare forecasts.

We note that, on a balanced dataset for which the event rate is 0.5, it can be shown that $\mathrm{TSS} = \mathrm{HSS} = 1-2(1-\mathrm{ACC})$. The trend and the paired t-test results for TSS apply to ACC and HSS due to perfect correlation. Therefore, we mainly focus on the discussion on TSS, list ACC as a complement metric, and omit HSS as it is equal to TSS in our setting. For probabilistic forecasts, the aforementioned metrics (ACC, HSS, and TSS) depend upon the threshold applied to the predicted probability. A common practice is to apply a threshold of 0.5, which is considered to be "random" by many researchers. In contrast, the following two metrics, BSS and AUC, are irrelevant to the threshold, and they need information (i.e., predicted probabilities) beyond the mere contingency table.

The Brier Skill Score (BSS) is a skill score evaluating the quality of a probability forecast. It is of a nature different from those of HSS and TSS, in that it directly uses probabilistic predictions without thresholding them. The BSS also admits the general skill score formulation, with the accuracy replaced by the Brier Score (BS), defined as the mean squared error between the probability predictions f_i's and binary outcomes o_i's:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2. \tag{13}$$

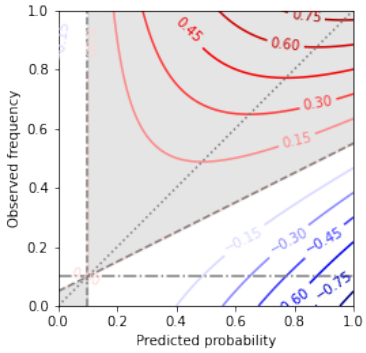


Figure 5. An illustration of the relation between the reliability diagram and BSS.

With a reference forecast that consistently predicts the average event frequency $\bar{o}$ (also known as climatology), the BSS is given by

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}_{\mathrm{forecast}}}{\mathrm{BS}_{\mathrm{reference}}} \,. \tag{14}$$

It is sometimes of interest to decompose BS into three components of reliability, resolution, and uncertainty (Murphy 1973; McCloskey et al. 2018). BSS is frequently accompanied by the reliability diagram discussed below, providing more complete information about the performance of probabilistic predictions.

For completeness we briefly discuss the Area Under Curve (AUC), defined as the area under the receiver operating characteristic (ROC) curve. The ROC curve depicts how the probability of detection changes with the false alarm rate by varying the classification threshold. A higher AUC implies a higher average detection probability over all false alarm rates. Unlike dichotomous metrics like TSS and HSS, AUC summarizes detection performance over all possible false positive rates and, in particular, does not depend on the threshold selected to convert probabilistic forecasts into binary decisions. Consequently, the AUC is not as useful in flare prediction, especially when stringent false positive control is exercised (Steward et al. 2017).

The above skill scores provide one way to directly compare flare prediction models. In addition to such metrics, flare forecasts are often evaluated using graphical tools for diagnostics and comparison. Apart from ROC curves discussed above, reliability diagrams (RD) and skill score profiles (SSP) are also commonly used in forecast verification. All three of them are applicable to forecasts that predict probabilities or continuous scores (e.g., logits) that can be converted to probabilities. We briefly discuss RD and SSP below.

The reliability diagram, also known as the calibration curve, measures how well a probabilistic forecast agrees with the observation. The predicted probabilities are binned into groups and the observed event rate within each group is plotted. If the predicted probability agrees well with the observed rate, the points will be close to the diagonal of the plot (the line of perfect reliability). Such a forecast is called reliable. Any forecast that produces predictions independent of flare activity has all its points close to the horizontal line at the event rate. BSS provides a metric that accounts for both reliability and resolution. Figure 5 shows an example of the plane on which the reliability diagram is drawn. The climatology rate is set to be 0.1. The overall BSS can be seen as a histogram weighted average of the contributions of the points on the reliability diagram. The contours are equal contribution lines. The points in the shaded area contribute positively to BSS. The dashed line with slope 1/2 is called the "no skill" line, the points on which have zero contribution to the overall BSS.

A skill score profile plot shows how skill scores change as a function of the probability threshold. A method with high and flat profile is usually desired, as such a method achieves high skill scores and the performance is robust to the changes of the threshold.

3.4. *Statistical performance comparisons*

We use a one-sided paired t-test to assess the comparative performance of a pair of prediction algorithms, called algorithm 1 and 2, that are tested on the same test data. Specifically, two competing hypotheses are formulated: the null hypothesis (H_0) that algorithms 1 and 2 have identical performance and the alternative hypothesis (H_1) that

algorithm 2 is better than algorithm 1. Suppose we have n pairs of empirical performance $\{(x_i, y_i)\}_{i=1}^n$ achieved by algorithm 1 and 2 on n test samples. The paired t-statistic is $t = \sqrt{n}(\overline{y} - \overline{x})/\sigma$ where $\overline{y}$ and $\overline{x}$ are the sample means of $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, respectively, and σ^2 is the sample variance of the differences $\{y_i - x_i\}_{i=1}^n$. Under H_0, t follows a Student-t distribution with $n-1$ degrees of freedom (Bickel & Doksum 2015). The p-value associated with the test statistic t is defined as the area under the Student-t density to the right of the value t. Small p-values provide strong evidence in favor of H_1, i.e., that algorithm 2 is better than algorithm 1. In this paper, we use the paired t-test to examine the following hypotheses:

- Training a predictor (LSTM or CNN) on data from two solar cycles (SMARP and SHARP) improves upon training a predictor on data from a single solar cycle (only SMARP or only SHARP) (Section 4.1).
- LSTM achieves better performance than CNN (Section 4.2).
- The LSTM-CNN stacking ensemble achieves better performance than the better model between LSTM and CNN (Section 4.3).

3.5. *Interpretation of CNNs*

Deep learning methods are widely applied to many domains such as computer vision, natural language processing, speech processing, robotics, and games (see, e.g. He et al. 2016; Silver et al. 2016; Devlin et al. 2018). As of today, deep learning algorithms largely remain black box methods, raising concerns of lack of interpretability, transparency, accountability, and reliability. Interpretability is of particular importance when deep learning is used in scientific discovery. Over the years, many tools for interpreting the functioning of deep neural networks have been proposed, revealing aspects of their underlying decision process.

One way to interpret a black-box model, often referred to as "attribution", is to see how different parts of the input contribute to the model's output. An attribution method generates a vector of the same size as the input, with each element indicating how much the corresponding element in the input contributes to the model decision for that input. In the context of CNNs, the attribution vector is a heatmap of the same size as the input image.

A multitude of attribution methods have been proposed for CNNs in the task of image classification. One type of approach is perturbation-based methods, among which occlusion (Zeiler & Fergus 2014) is well known. Occlusion masks the input image with a gray patch at different locations and sees how much the prediction score of the ground truth class drops. The prediction score drop varies with location, forming a heatmap, with large values indicating the positions of the features important to the CNN's correct prediction. One drawback of the occlusion method is that it is computationally expensive. Another drawback is that the attribution depends on the size and shape of the patch, which need to be tuned to obtain sensible results. Therefore, this type of approach is not used in our work.

Another type of approach uses gradient-based methods, the basic idea being that the gradient of the predicted score of a certain class with respect to the input reveals the contribution of each dimension of the input. Saliency map (Simonyan et al. 2013), one of the earliest gradient-based methods, is simply the absolute value of the gradients. The intuition is that the magnitude of the derivative indicates which pixels need to be changed the least to affect the class score the most (Simonyan et al. 2013). Deconvolution Network (Zeiler & Fergus 2014) and Guided Backpropagation (Springenberg et al. 2015) attribution methods modify the backpropagation rule. Integrated Gradients (Sundararajan et al. 2017) integrate the gradients along the path from a reference image to the target image. Formally, the integrated gradient along the i-th dimension for an input x and a baseline x' is

$$L_i^c(x; x') = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F_c(x' + \alpha \times (x - x'))}{\partial x_i} \, \mathrm{d}\alpha, \tag{15}$$

where $F_c(x)$ is the model output for class c with input x. One desirable property of Integrated Gradients, known as completeness, is that the pixels in the attribution map add up to the difference of prediction scores of the target and the reference image, i.e., $F(x) - F(x')$. DeepLIFT (Shrikumar et al. 2017) and its gradient-based interpretation (Ancona et al. 2018) can be seen as the gradient with modified partial derivatives of non-linear activations with respect to their inputs. Grad-CAM (Gradient-weighted Class Activation Mapping) (Selvaraju et al. 2017) accredits decision-relevant signatures by generating a saliency map, highlighting pixels in the input image that increase the confidence of the network's decision for a particular class. More formally, the Grad-CAM heatmap L^c for class c with respect to a

particular convolutional layer is given by the positive part of the weighted sums of the layer's activation maps A_k, i.e.,

$$L^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right), \tag{16}$$

with weights α_k^c given by the spatial average of partial derivatives of the class-specific score y^c with respect to the class activation map as

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}, \tag{17}$$

where Z is a normalization constant. Intuitively, a class activation map is weighted more if the pixels therein make the CNN more confident in its decision that the input belongs to class c.

In solar flare prediction, Bhattacharjee et al. (2020) applied the occlusion method and found that CNNs pay attention to polarity inversion regions. Yi et al. (2021) applied Grad-CAM to CNNs and found that polarity inversion lines in full-disk MDI and HMI magnetograms are highlighted as an important feature for flare prediction. In this paper, using a variety of attribution methods, we observe similar trends for the CNN trained on SHARP and SMARP data.

4. RESULTS

We aim to answer the following four scientific questions: (1) Can additional data from another solar cycle benefit the performance of deep learning methods for solar flare prediction? (2) Do features implicitly learned by CNN work better than handcrafted physical parameters used by LSTM? (3) Can we combine the two deep learning methods to obtain a better prediction? (4) What preflare signatures can the CNN detect from the magnetogram of an active region?

To summarize, we report the following findings: (1) Additional training data from HMI collected in Solar Cycle 24 improve the predictive performance of both LSTM and CNN when tested on Solar Cycle 23. (2) LSTM (using summary parameters) generally outperforms CNN (directly using magnetograms) in flare prediction. (3) Stacking CNN and LSTM generally leads to better prediction performance. (4) Visual attribution methods help us interpret the decision of CNN by identifying preflare features. This section presents the empirical results that lead to these findings.

4.1. *Data from another solar cycle improves prediction*

A major goal of this paper is to examine the utility of using SMARP and SHARP together. We set an experimental group and a control group and contrast their 24-hour "strong-vs-quiet" flare prediction performance. The control group consists of models that train, validate, and test exclusively on SHARP data. We refer to this type of dataset as `SHARP_ONLY`. Compared to the control group, models in the experimental group have the training set enriched by SMARP data, while the validation and the test set are kept the same. We call this type of dataset `FUSED_SHARP`. The only difference between `SHARP_ONLY` and `FUSED_SHARP` is that models using `FUSED_SHARP` have access to data from a previous solar cycle in the training phase. Symmetrically, we design `SMARP_ONLY` and `FUSED_SMARP` to examine the utility that SHARP brought to SMARP. Specifically, the four types of datasets are generated as follows:

1. Dataset `SHARP_ONLY`: 20% of all the HARPs are randomly selected to form a test set. 20% of the remaining HARPs are randomly selected to form a validation set. The rest of the HARPs belong to the training set. In each split, negative samples are randomly selected to match the number of positive samples.

2. Dataset `FUSED_SHARP`: The test set and the validation set stay the same, respectively, with those in `SHARP_ONLY`. The remaining HARPs are combined with all TARPs to form the training set. In each split, negative samples are randomly selected to match the number of positive samples.

3. Dataset `SMARP_ONLY`: 20% of all the TARPs are randomly selected to form a test set. 20% of the remaining TARPs are randomly selected to form a validation set. The rest of the TARPs belong to the training set. In each split, negative samples are randomly selected to match the number of positive samples.

4. Dataset `FUSED_SMARP`: The test set and the validation set stay the same, respectively, with those in `SMARP_ONLY`. The remaining TARPs are combined with all HARPs to form the training set. In each split, negative samples are randomly selected to match the number of positive samples.

Table 5. Sample sizes of a random realization of the four datasets

	Train		Validation		Test	
	Positive	Negative	Positive	Negative	Positive	Negative
SHARP_ONLY	1774	1774	665	665	410	410
FUSED_SHARP	6375	6375	665	665	410	410
SMARP_ONLY	2849	2849	860	860	892	892
FUSED_SMARP	5698	5698	860	860	892	892

Table 6. Test set performance of the LSTM and the CNN on 24-hour "strong-vs-quiet" flare prediction. The two datasets within each comparison group share common test sets. The 1-σ error is calculated from 10 random experiments. Bold fonts indicate the experiments in which the mean of the metric on the fused dataset is higher than that on the single dataset.

		Group 1		Group 2	
	Dataset / Model	FUSED_SHARP	SHARP_ONLY	FUSED_SMARP	SMARP_ONLY
ACC	CNN	0.906+/-0.036	0.922+/-0.017	**0.901+/-0.028**	0.877+/-0.031
	LSTM	**0.950+/-0.012**	0.942+/-0.016	**0.905+/-0.025**	0.900+/-0.024
AUC	CNN	0.980+/-0.009	0.981+/-0.006	**0.963+/-0.017**	0.950+/-0.020
	LSTM	**0.990+/-0.004**	0.986+/-0.004	**0.966+/-0.015**	0.963+/-0.015
TSS	CNN	0.812+/-0.071	0.843+/-0.034	**0.802+/-0.056**	0.754+/-0.061
	LSTM	**0.900+/-0.023**	0.884+/-0.032	**0.810+/-0.050**	0.800+/-0.049
BSS	CNN	0.649+/-0.152	0.714+/-0.064	**0.628+/-0.114**	0.520+/-0.121
	LSTM	**0.799+/-0.036**	0.775+/-0.047	**0.626+/-0.107**	0.586+/-0.108

Since train/validation/test split and undersampling are both random, repeating these two steps with different seeds enables uncertainty quantification to the evaluation results. The tally of samples produced by one particular random seed is shown in Table 5. On each of the four types of datasets, LSTMs and CNNs are fitted on the training set, validated on the validation set, and evaluated on the test set.

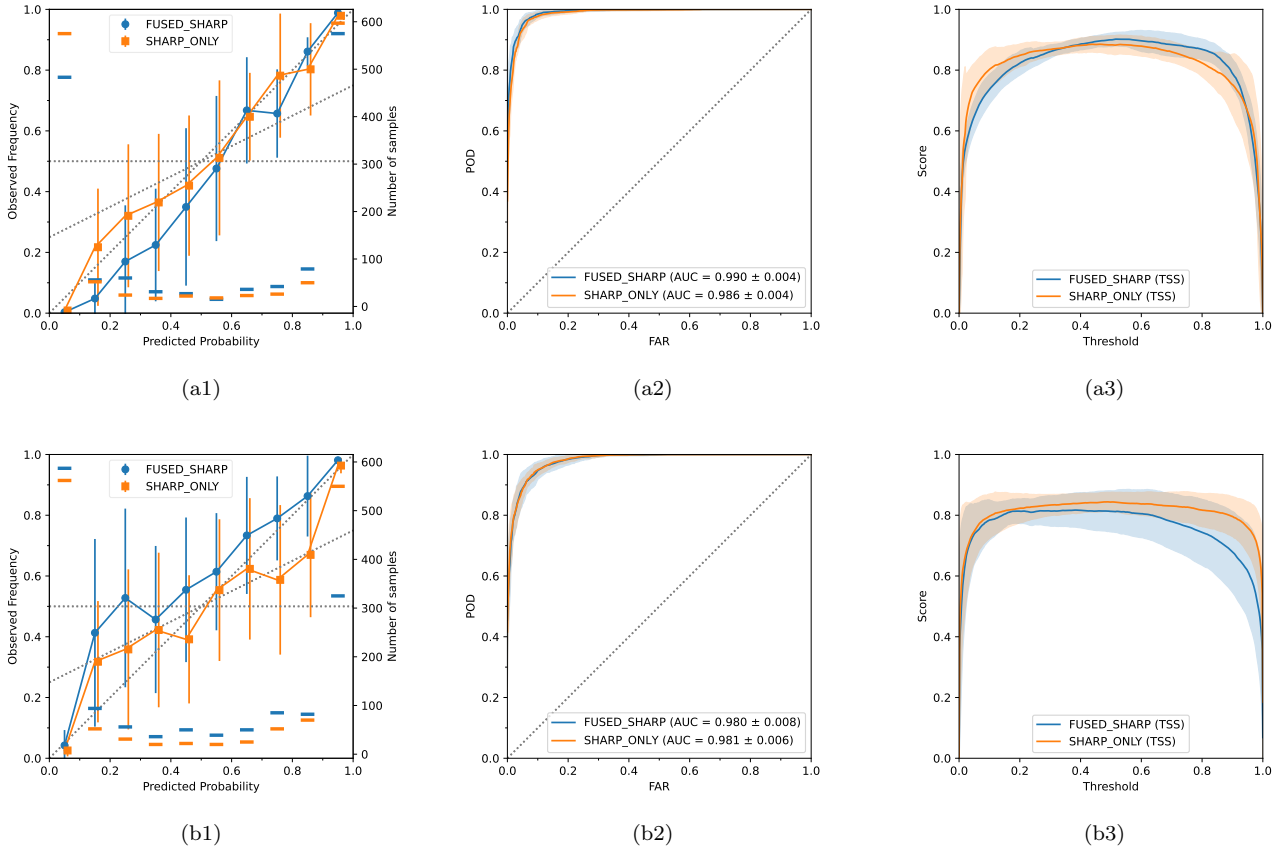
Table 6 shows the results of the "strong-vs-quiet" active region prediction using the LSTM and the CNN. For LSTMs, a consistent improvement on the fused datasets (FUSED_SHARP and FUSED_SMARP) is observed in terms of the mean of all metrics. This aligns with the fact that more data are typically desired to improve the generalization performance of deep learning models because they are overparameterized and can easily overfit on small datasets. For CNNs, an improvement is observed on FUSED_SMARP over SMARP_ONLY, but not on FUSED_SHARP over SHARP_ONLY. This indicates that the lower image quality in SMARP has a negative effect on CNN's performance.

The statistical significance of the improvement on the fused datasets is tested using a one-sided paired t-test with significance level 95%. Table 7 shows the t-statistics and the associated p-values of the paired t-tests. The bold font p-values are less than 0.05 and considered to be significant. For LSTMs, the fused datasets are better than the single datasets in a statistically significant way in almost all settings. The only exception is BSS on FUSED_SHARP, whose p-value is only slightly larger than 0.05. For CNNs, across all metrics, statistically significant improvement is observed for FUSED_SMARP over SMARP_ONLY, but not for FUSED_SHARP over SHARP_ONLY. This indicates that adding SHARP magnetograms into SMARP during training helps the CNN to better predict flares, but not the other way around. One potential reason is SMARP magnetograms have a lower signal-to-noise ratio than SHARP magnetograms, which may have negatively affected the CNN. The LSTM, on the other hand, uses the active region summary parameters, which could suppress the effect of noise during summarizing magnetograms, providing information in a sufficiently good quality that does not offset the improvement induced by the increased training sample size.

Aside from the numerical metrics, we provide graphical evaluation results for Group 1 (FUSED_SHARP and SHARP_ONLY) in Figure 6, and Group 2 (FUSED_SMARP and SMARP_ONLY) in Figure 7. A trend of over-forecasting for high probabilities

Table 7. Paired t-tests for significant improvement of test set performance on the fused datasets as measured by different metrics. The alternative hypothesis H_1 claims that metric S on the fused dataset (`FUSED_SHARP` or `FUSED_SMARP`) is greater than the respective single dataset (`SHARP_ONLY` or `SMARP_ONLY`), which is tested against the null hypothesis H_0 claiming otherwise. The bold font p-values are less than 0.05 and considered to be significant.

	H_1	$S_{\texttt{FUSED\_SHARP}} > S_{\texttt{SHARP\_ONLY}}$		$S_{\texttt{FUSED\_SMARP}} > S_{\texttt{SMARP\_ONLY}}$	
		p-value	t	p-value	t
Metric S	Model				
ACC	CNN	0.885787	-1.292359	**0.001862**	3.881137
	LSTM	**0.016544**	2.514074	**0.026797**	2.219666
AUC	CNN	0.589845	-0.233881	**0.001399**	4.070352
	LSTM	**0.000459**	4.842485	**0.033930**	2.074572
TSS	CNN	0.885787	-1.292357	**0.001862**	3.881135
	LSTM	**0.016544**	2.514079	**0.026796**	2.219673
BSS	CNN	0.889419	-1.314583	**0.000482**	4.806837
	LSTM	0.054812	1.775082	**0.000099**	6.014784



(a1) (a2) (a3)

(b1) (b2) (b3)

Figure 6. Verification plots on SHARP test data to compare models trained on `FUSED_SHARP` and `SHARP_ONLY`. Shown in (a1)–(a3) are the reliability diagram, ROC, and SSP for LSTM. Shown in (b1)–(b3) are the same plots but for CNN. In each panel, the blue/orange curve is the test performance for the model trained on `FUSED_SHARP`/`SHARP_ONLY`. In each graph, solid curves and error bars (or shaded area) indicate respectively the means and the standard deviations calculated from 10 random experiments. In each reliability plot, the short horizontal bars indicate the number of samples in each probability bin, and the two curves are separated horizontally to prevent error bars from overlapping.

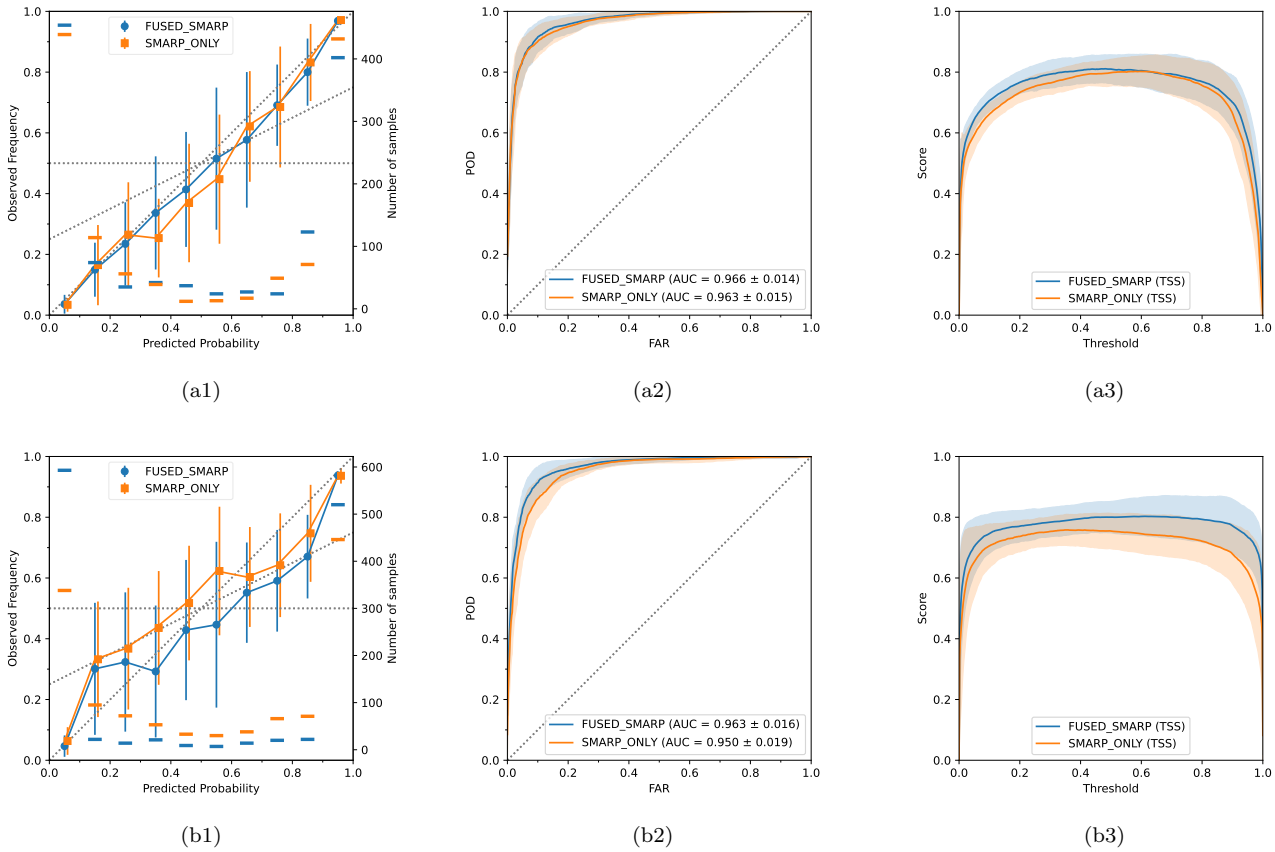


Figure 7. Same as Figure 6 but for SMARP test data to compare models trained on `FUSED_SMARP` and `SMARP_ONLY`.

and under-forecasting for low probabilities is observed in some cases but such effect is minor considering the size of the error bars. In reliability diagrams, all models have points closer to the diagonal, indicating high reliability. In ROC plots, it is observed that the LSTM achieves higher AUC on the fused datasets (`FUSED_SHARP` and `FUSED_SMARP`) than on the single datasets (`SHARP_ONLY` and `SMARP_ONLY`). For the CNN, similar improvement is also observed in the comparison of `FUSED_SMARP` and `FUSED_SHARP`, whereas the ROCs are almost indistinguishable for `FUSED_SHARP` and `SHARP_ONLY`. In skill score profiles, the TSS for LSTM trained on fused datasets are at the same level as that trained on single datasets. For the CNN, on the other hand, `FUSED_SHARP` displays a disadvantage against `SHARP_ONLY`, whereas `FUSED_SMARP` displays an advantage over `SMARP_ONLY`. This verifies the observations made from metrics. In all cases, the skill score profiles are high and relatively flat, indicating the robustness of the performance to the change of thresholds within a wide range of the varying threshold.

4.2. *LSTM performs better than CNN*

This section provides forecast verification to the LSTM and the CNN. We use the same evaluation results for 10 experiments in each setting mentioned in Section 4.1, but present them in a way that makes it easier to compare the LSTM and the CNN. We note the differences between our verification set-up and that in an operational setting:

1. In terms of data, the test set of our sort has lots of samples removed based on their active regions, observational data, and flare activities. About 1/5 of tracked active region time series in the evaluation period (May 2010–December 2020) are selected. Within each active region series, only samples with good quality observation and certain flaring patterns are selected (detailed in Section 2.3). Negative samples (flare-quiet active regions) are significantly downsampled to match the number of positive samples (strong-flare-imminent active regions). In contrast, operational forecasts do not discard any sample unless absolutely necessary.

Table 8. Paired t-tests for significant improvement of the LSTM over the CNN in terms of different metrics S on the test set of the four datasets. The alternative hypothesis H_1 claims $S_{\text{LSTM}} > S_{\text{CNN}}$. The bold font p-values are less than 0.01 and considered to be significant.

Dataset	FUSED_SHARP		SHARP_ONLY		FUSED_SMARP		SMARP_ONLY	
Metric S	p-value	t	p-value	t	p-value	t	p-value	t
ACC	**0.001442**	4.050296	**0.007142**	3.028403	0.234866	0.754672	**0.001079**	4.245351
AUC	**0.003757**	3.429557	**0.002527**	3.682754	0.227978	0.779031	**0.000743**	4.501005
TSS	**0.001442**	4.050297	**0.007142**	3.028405	0.234865	0.754673	**0.001079**	4.245350
BSS	**0.005296**	3.213872	**0.002645**	3.653351	0.531965	-0.082481	**0.005781**	3.159315

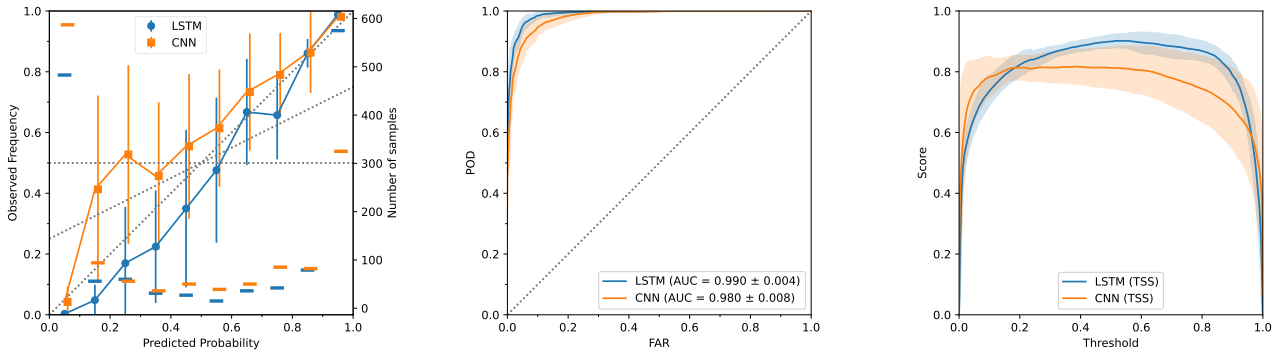


Figure 8. Verification plots of the LSTM and the CNN on FUSED_SHARP. Shown are the reliability diagram, ROC, and SSP, from left to right. This figure essentially extracts the blue curves (representing FUSED_SHARP) in both rows of Figure 6 and overlaps them together.

2. In terms of outcomes, the forecast of our sort is independent for individual active regions, with the prediction result available every 96 minutes (i.e., MDI cadence) for valid active regions. In contrast, the end goal of an operational forecast is a full-disk forecast. For operational forecasts built upon active region based forecasts, the predictions for all active regions on the solar disk are aggregated to compute the full-disk prediction. In addition, operational forecasts are typically issued at a lower frequency (e.g., every 6 hours), but in a consistent manner.

The verification results in this section should be interpreted with the above differences in mind.

It can be seen from Table 6 that the LSTM generally scores higher than the CNN in terms of mean performance. We performed paired t-test to validate this observation. The results in Table 8 confirm that the LSTM scores significantly higher ($p < 0.01$) than the CNN across all metrics on all datasets except for FUSED_SMARP. On FUSED_SMARP, although we cannot claim statistical significance, the LSTM's performance is slightly better or at the same level as the CNN as is observed from Table 6.

We only present the graphical verification results for both models trained and tested FUSED_SHARP, given that SHARP is widely used and validated by a wealth of studies. For the results on other datasets, the visualization can be obtained by simply rearranging the same results shown in Figure 6 and 7.

The reliability diagram in Figure 8 shows that the probabilistic prediction given by the LSTM is closer to the diagonal than the CNN, and hence more reliable. The CNN exhibits a trend of under-forecasting especially when the predicted probability is less than 0.5. The histogram of predicted probability shows that probabilistic forecast by the LSTM is "more confident", or has higher resolution, than LSTM, with most of the predicted probabilities close to 0 or 1.

The ROC in Figure 8 shows a clear advantage of the LSTM over the CNN, in the sense that it achieves a higher probability of detection with the same false alarm rate. This trend is also manifested in terms of AUC.

The SSP in Figure 8 shows LSTM achieves higher TSS on average for all thresholds within 0.2–0.9. It is also observed that the TSS for LSTM is maximized by a threshold very close to the climatological rate on the test set (which is 0.5 in our case), a necessary condition for a reliable predictor (Kubo 2019).

4.3. *Stacking LSTM and CNN leads to better prediction*

In this paper, we only consider stacking methods to combine the CNN and the LSTM hoping for better predictive performance. We evaluate the test set performance of stacking methods using four different criteria:

- `CROSS_ENTROPY`: weights are optimized to minimize cross-entropy loss on the validation set.
- `BSS`: weights are optimized to maximize BSS on the validation set.
- `AUC`: weights are optimized to maximize AUC on the validation set.
- `TSS`: weights are optimized to maximize TSS on the validation set.

Among these criteria, cross-entropy and negative BSS are known to be convex; TSS is neither convex nor concave; we observe AUC to be concave but we do not have proof other than empirical evidence. Criteria `HSS` and `ACC` are excluded from the evaluation since their stacking weights are the same as that of `TSS` due to the perfect correlation mentioned in Section 3.3.

To provide baseline performances, we include the evaluation results for the two base learners, `LSTM` and `CNN`. In addition to the above stacking methods, we consider two other meta-learning schemes:

- `AVG` outputs the average of predicted probabilities of two base learners.
- `BEST` (Džeroski & Ženko 2004) selects the base learner that performs the best on the validation set and applies it to the test set.

Splitting and undersampling are randomly performed 10 times on each of the four datasets `FUSED_SHARP`, `FUSED_SMARP`, `SHARP_ONLY`, and `SMARP_ONLY`. The test set TSS of the 10 random experiments for each criterion on each dataset are summarized as box plots in Figure 9. The optimal stacking weights for the four stacking ensembles are summarized in Figure 10.

Figure 9 shows that stacking methods perform slightly better than the `BEST` meta-learner, especially on `FUSED_SMARP` and `SMARP_ONLY`. Of note, the wide error bars are partially due to the randomness originating from data sampling. To fairly compare the methods, we perform paired t-tests with significance level 0.05. It turned out stacking is significantly better than `BEST` in the following three settings: `BSS` on `FUSED_SMARP` ($p = 0.048$), `AUC` on `SMARP_ONLY` ($p = 0.025$), and `TSS` on `SMARP_ONLY` ($p = 0.013$).

We also note in Figure 9 that `BEST` unsurprisingly achieves better performance than `AVG` but is slightly worse than the better performing base learner `LSTM`, most noticeably on `FUSED_SHARP`. In fact, `BEST` decided that `CNN` is the better model in 3 out of 10 experiments on `FUSED_SHARP`. This is not unexpected because the "best" model on the validation set is not necessarily the best on the test set.

From Figure 10, we can see that α is greater than 0.5 in most experiments, with the median falling between 0.55 and 0.9 in all settings. This suggests that stacking ensembles generally depend more on the LSTM than on the CNN. The variance of α is large in some settings, especially for the `AUC` on `FUSED_SMARP`. The variance of convex criteria (`CROSS_ENTROPY` and `BSS`) is not smaller than that of nonconvex criteria (`TSS`), indicating that the local minima of non-convex loss functions is not the major source of variance. We suspect the major source of the variance comes from the data sampling bias among experiments, which is, in turn, a collective consequence of the insufficient sample size, heterogeneity across active regions, and possibly a small amount of information leakage because the validation set is used both in the validation of base learners and the training of the meta-learner.

We inspect one experiment of stacking with criterion `ACC` and the results are presented in Figure 11. Figure 11 (a1)–(a3) show the predicted probabilities by the LSTM and the CNN of each instance in the training, the validation, and the test set. The points are colored by their labels, with red representing the positive class and blue representing the negative class. The green solid line in (a2) and (a3) shows the decision boundary by the meta-learner with α fitted on the validation set to maximize ACC. The points (p, q) on the upper right side of the boundary are classified as positive because they satisfy $r = \alpha p + (1 - \alpha)q > 0.5$. In this experiment, the fitted $\alpha = 0.586$, suggesting the stacking

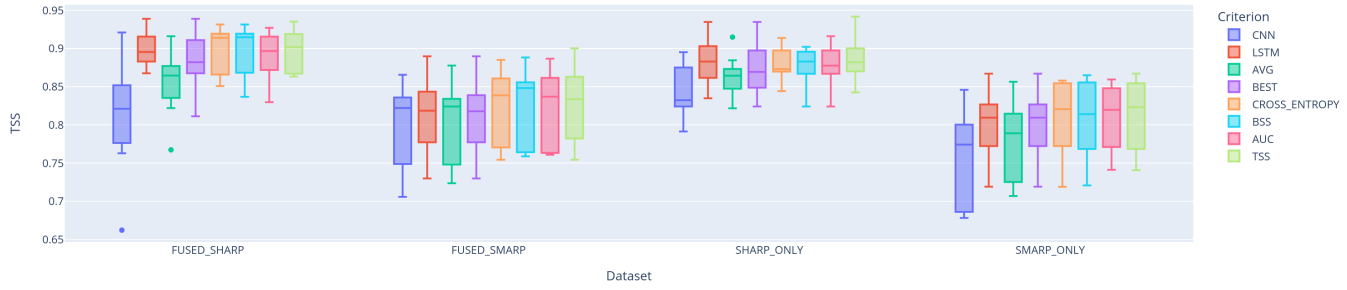


Figure 9. Test set TSS for base learners and meta-learners using different criteria.

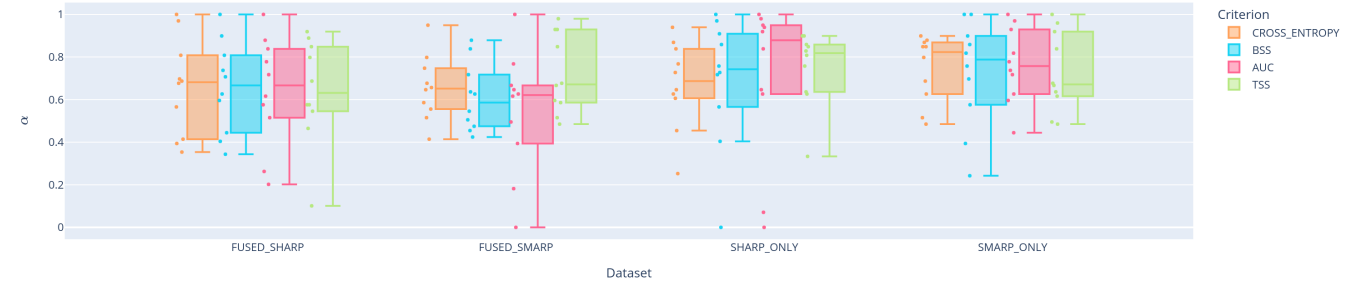


Figure 10. Stacking weight α fitted using different criteria on different datasets. All 10 values of α in an experiment setting are shown as points next to the corresponding box.

ensemble relies almost equally on the CNN and the LSTM. The violet dashed line in (a3) is the decision boundary with α fitted on the test set, and hence can be seen as the oracle. It can be observed that the distribution of predicted probabilities on the validation set (a2) and the test set (a3) are similar. The distribution of predicted probabilities on the training data in (a1), on the other hand, looks completely different, with the CNN achieving almost perfect separation. In fact, the CNN overfitted on the in-sample data, as indicated by a significantly lower positive recall rate in (a2) and (a3). This validates the decision that meta-learners should not be fitted on the predicted probabilities of the same data used to train the base learners.

Figure 11(b) exhibits the stacking optimization process for the same experiment, in which the ACC is calculated on the validation set (a2) and the test set (a3) by scanning over a fine grid of $\alpha \in [0, 1]$ with resolution 0.001. Although the validation ACC (blue curve) is not concave with respect to α, it does exhibits a maximum at $\alpha = 0.586$ as indicated by the vertical green line. The stacking ensemble's test set performance over α is shown as the red curve. These curves indicate that stacking the LSTM and the CNN indeed results in a small improvement of performance relative to implementing either of them alone, corresponding to the values of ACC at $\alpha = 1$ or $\alpha = 0$.

4.4. *CNN identifies the emergence of preflare features*

We use visual attribution methods to extract flare-indicative characteristics of magnetograms from trained CNNs. First, we use synthetic images to examine patterns that contribute to a positive decision of CNNs. The results of synthetic images help us understand better the attribution maps of real magnetograms. Then, we apply visual attribution methods to image sequences of selected active regions that transition from a flare-quiescent state to a flare-imminent state. Setting the baseline to the first image in the sequence gives a time-varying attribution map that tracks magnetic field variations that contribute to the change in the predicted probability.

4.4.1. *Synthetic image*

To assist our understanding of attribution maps obtained by different methods, we first turned to synthetic magnetograms. We take the bipolar magnetic region (BMR) model in Yeates (2020), represented as line-of-sight magnetic

(a1) Train set (a2) Validation set (a3) Test set (b) Optimization of ACC

Figure 11. (a1)–(a3): CNN predicted probability (y-axis) vs. LSTM predicted probability (x-axis) for the train, the validation, and the test set. The green solid line in (a2) and (a3) is the decision boundary of the ensemble with meta-learner fitted on the validation set. The violet dashed line in (a3) is analogous to the green line except that it is fitted on the test set, and hence can be seen as the oracle. (b): ACC as a function of α on the validation and the test set. The vertical green line shows the value of α that maximizes the validation ACC. The leftmost values of the ACC curves ($\alpha = 0$) correspond to the ACC of the CNN, and the rightmost values of these curves ($\alpha = 1$) correspond to the ACC of the LSTM.

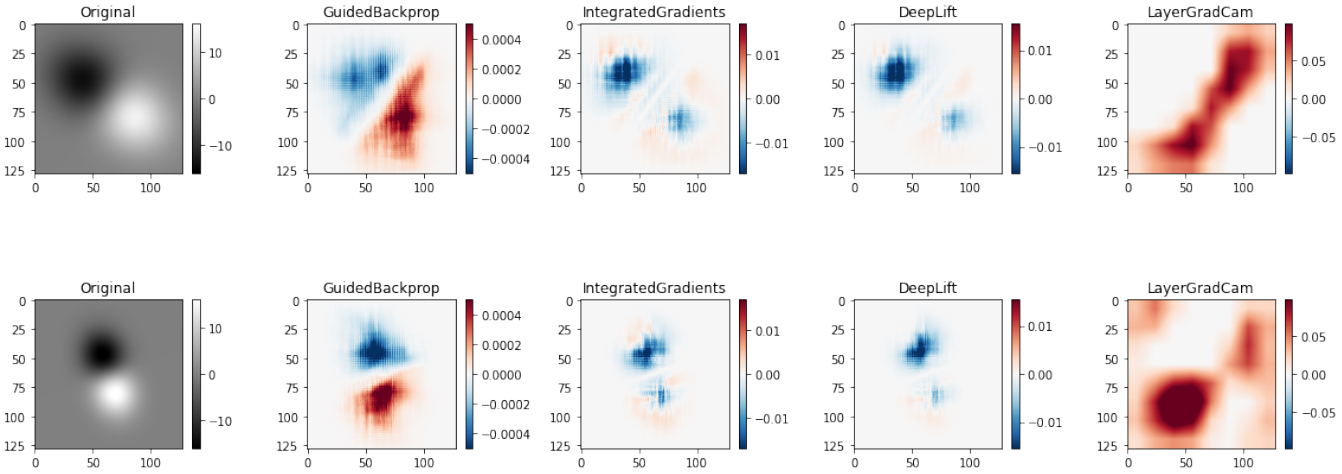


Figure 12. Examples of synthetic bipole images and attribution maps.

field B as a function of Heliographical location (s, ϕ), where s denotes sine-latitude and ϕ denotes Carrington longitude. B is parameterized by amplitude B_0, polarity separation ρ (in radian), tilt angle γ (in radian) with respect to the equator, and size factor a fixed to be 0.56 to match the axial dipole moment of SHARP (Yeates 2020). The untilted BMR centered at origin has the form

$$B(s, \phi) = -B_0 \frac{\phi}{\rho} \exp\left[-\frac{\phi^2 + 2\arcsin^2(s)}{(a\rho)^2}\right]. \tag{18}$$

We sweep a grid of B_0, ρ, and tilt angle γ to generate a BMR dataset. Of particular interest are synthetic BMRs considered to be flare-imminent by CNNs. Figure 12 shows some examples of them and their attribution results, from which patterns of positive predictions can be summarized. Guided Backpropagation heatmaps have both poles highlighted with the signs matching the polarities. Integrated Gradients produces heatmaps that are more concentrated to polarity centers and attribute more credits to the negative polarities. DeepLIFT produces similar heatmaps to those by Integrated Gradients. Grad-CAM's results are not as interpretable as the above methods. They seem to avoid the polarities and highlight the background and sometimes the polarity inversion lines.

4.4.2. *The emergence of preflare signatures in the active region evolution*

We focus on the attribution results on SHARP as opposed to SMARP because the former has magnetograms of higher resolution and lower noise level. We choose the CNNs that are trained on `SHARP_ONLY` as opposed to `FUSED_SHARP`

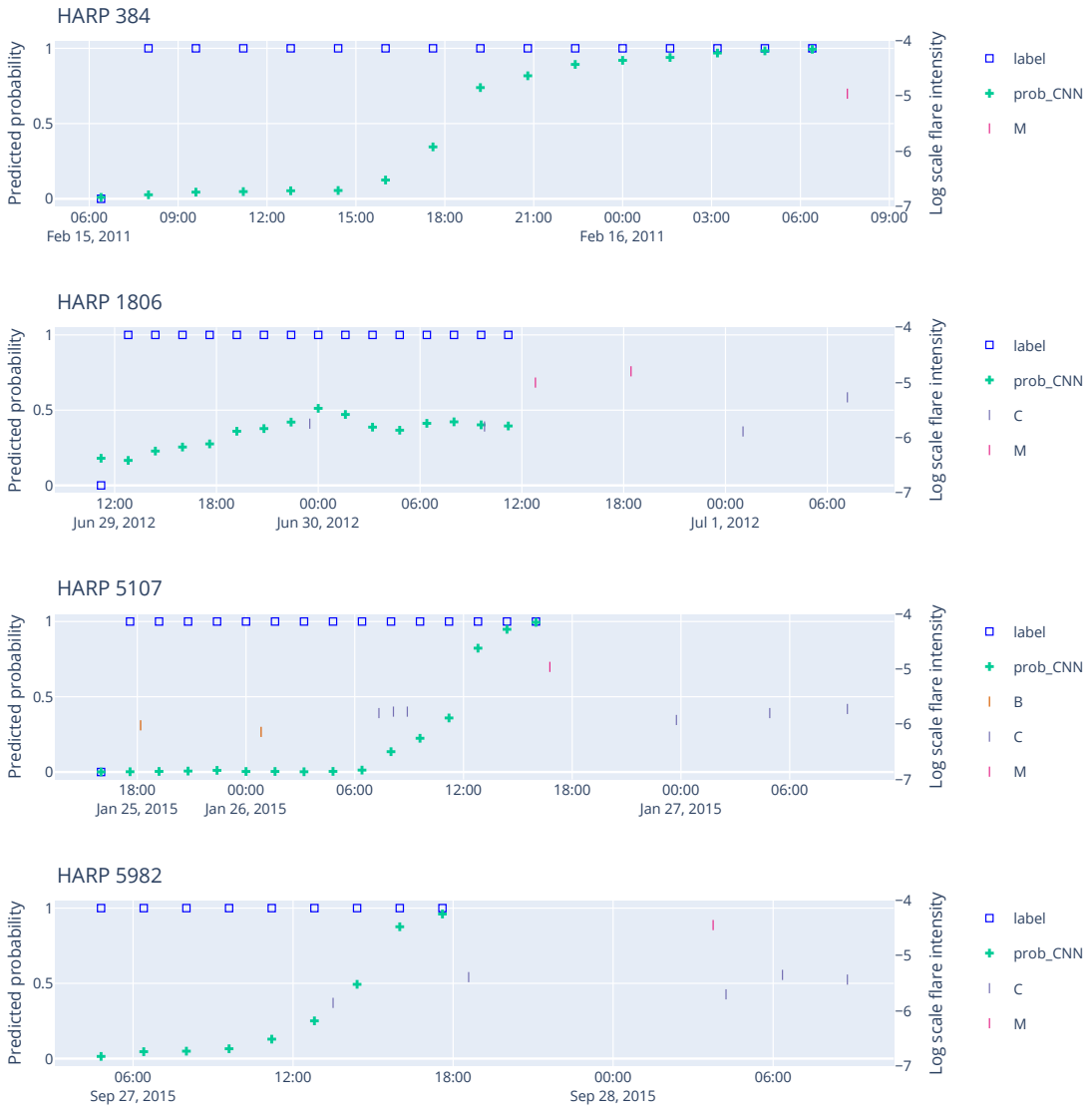


Figure 13. CNN predictions of part of time series of in HARP 384, 1806, 5107, and 5982. The labels are shown as blue open boxes and predicted probabilities as green plus symbols. The point-in-time instance is labeled as positive if an M1.0+ flare occurred in the future 24 hours in that active region. GOES flare events during and 24 hours within the sample sequence are shown as short vertical bars, with y-coordinates indicating flare intensities (peak flux in W/m^2) on a log scale.

because the former is observed to generalize better according to Section 2. To get results that reflect the generalization performance as opposed to training artifacts, we need to make sure that active regions being investigated are out-of-sample. To evaluate any active region of interest in SHARP, we perform 5-fold cross-validation on `SHARP_ONLY`, so that every active region is associated with a CNN that has never seen the active region in training. In addition, we do not enforce the flare-based sample selection rules and random undersampling, so that the evolution of attribution maps can be evaluated more coherently. As case studies, we select four HARP sequences that transition from a flare-quiescent state to a flare-imminent state. Figure 13 shows the labels and predicted probabilities of the four sample sequences. The attribution methods are performed on each HARP sequence in a frame-by-frame manner.

Figure 14 shows the last image of the four HARP sample sequences. The attribution maps of the same size as the input of the CNN (128× 128 pixels) are upsampled to the original resolution of the SHARP magnetogram using the `resize` method of the Python package `skimage.transform` with 2nd-order spline interpolation. The attribution maps

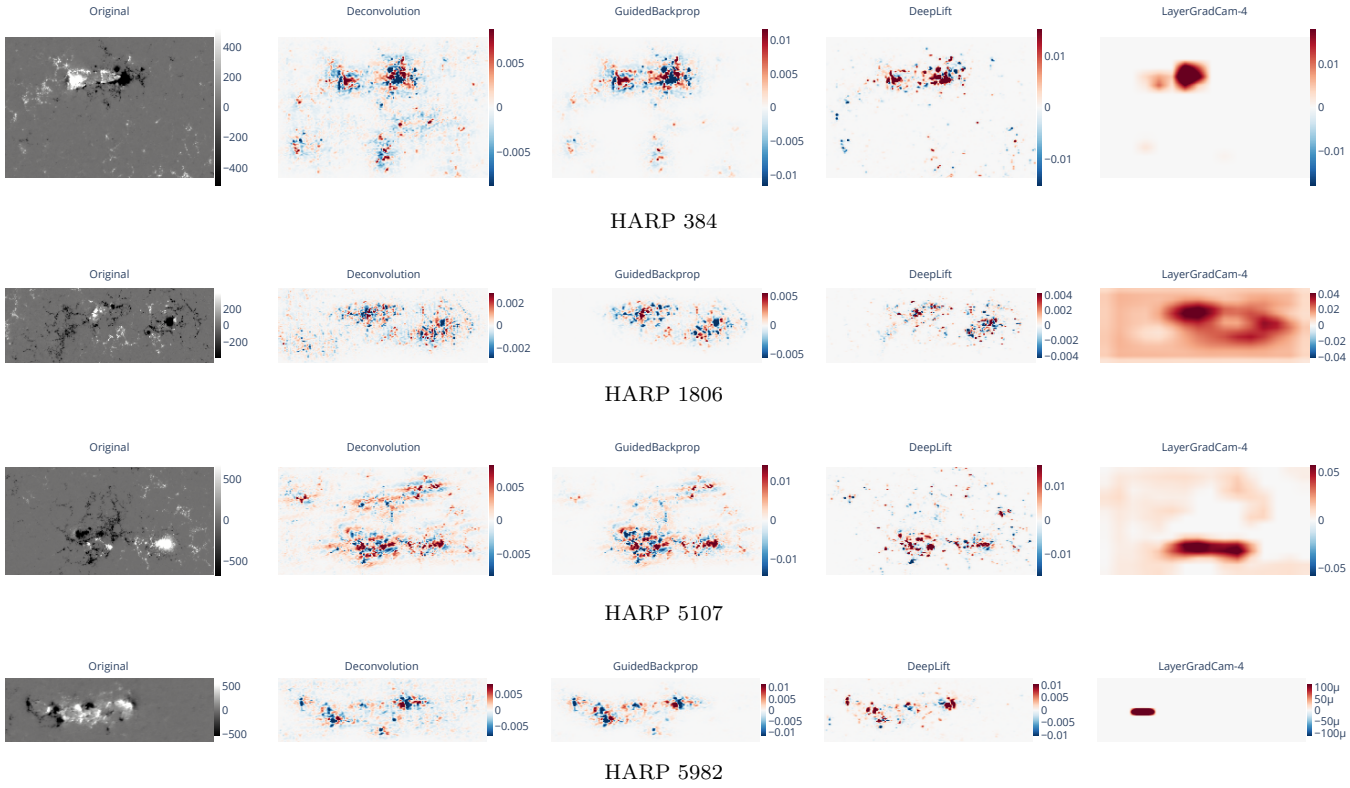


Figure 14. Attribution results of Deconvolution, Guided Backpropagation, DeepLIFT, and Grad-CAM on the last magnetogram in the sample sequences of HARP 384, 1806, 5107, and 5982. DeepLIFT chooses the first sample in the sequence as the reference. "LayerGradCam-4" means Grad-CAM with respect to the output of the fourth, or the second to last, convolutional layer. The interactive movie of heatmaps on all 9 samples in HARP 5982 using more attribution methods can be accessed at https://zeyusun.github.io/attribution/captum_movie_first.html.

of DeepLIFT and Integrated Gradients are similar. As such, only the results of the former are shown. The results for Integrated Gradients can be accessed online with the link shown in the caption.

In Figure 14, the attribution maps of Guided Backpropagation are observed to be more concentrated in strong fields compared to that of Deconvolution. The reference image of DeepLIFT and Integrated Gradients are chosen as the first sample in each sequence. From these two methods, the change of the prediction scores is attributed to the change of magnetic configuration of the last frame relative to the first frame, with red pixels indicating positive contribution and blue pixels indicating negative contribution. Since the predicted event probability of the last frame is higher than the first frame for all HARPs (Figure 13), the red pixels outweigh the blue pixels in the attribution maps of DeepLIFT and Integrated Gradients. The Grad-CAM results roughly reveal the position of the strong fields and polarity inversion lines.

From the visual attribution map, the CNN's prediction of a flaring active region can be accredited to the elements in the magnetogram. Figure 15 shows the contour plots of attribution maps generated by Integrated Gradients overlaid on magnetograms of the four HARP series. The contours enclose areas with large absolute values of Integrated Gradients in the last frame of each series, with red/blue contours indicating the region contributing positively/negatively to the increase in predicted probability. A general pattern is that the flux is emerging in red contours and canceling in blue contours. From the attribution maps, we can explain the increase in prediction scores as the consequence of the emerging flux outweighing the canceling flux.

The visual attribution maps can not only be used to identify preflare signatures in an active region; comparing them with our knowledge of flaring active regions can provide insights to diagnose, and potentially improve, the machine learning method used to predict flares. Here we provide an example. A known artifact in magnetograms is the fake polarity inversion line (PIL) caused by the projection effect when the magnetic vector's inclination relative to the line-of-sight surpasses 90° (Leka et al. 2017). In Figure 15(d), the emerging polarity inversion line in the penumbra of

the leading polarity (on the right/west part of the active region) is picked up as a preflare signature by the largest red contour. However, HARP 5982 is on the limb of the solar disk at the time (Figure 16), and the emerging PIL is caused by the highly inclined magnetic field in the penumbra as the flux rope is elevating from the surface. This shows that the CNN trained to associate magnetograms and flaring activities is not able to discern the polarity artifact by itself. This also suggests that the model could be potentially improved if we feed the location information to the CNN to help it correct such artifact. A similar PIL artifact is also observed in the following polarity of HARP 5107 in Figure 15(c). Since this artifact does not change much during the observation interval, it does not contribute as much to the change of the prediction score.

We remark the attribution maps obtained by Integrated Gradients are better in terms of resolution and interpretability than what were used in Bhattacharjee et al. (2020) and Yi et al. (2021). The occlusion method in Bhattacharjee et al. (2020) was shown to highlight the area between the opposite polarities, providing only crude attribution. This is because the size of the occlusion mask is usually chosen to be big enough to cover the informative regions. The result of Grad-CAM, being the attribution to a convolutional layer as opposed to the input, also suffers from the low-resolution issue. Both the Grad-CAM results in Figure 14 and in Yi et al. (2021) are able to highlight active regions, but the resolution is not high enough to reveal any structural information within the active region at the level of magnetic elements. Guided Backpropagation in Yi et al. (2021) is able to identify polarity inversion lines. However, it has been observed (and theoretically assessed) that Guided Backpropagation and Deconvolution behave similarly to an edge detector, i.e., they are activated by strong gradients in the image and insensitive to network decisions (e.g. Nie et al. 2018; Adebayo et al. 2018). In contrast, the method of Integrated Gradients needs a baseline, which aligns with the natural way in which the human interprets an observation: by assigning credit or blame to a certain cause, we implicitly consider the absence of the cause (Sundararajan et al. 2017). In addition, Integrated Gradients essentially "decomposes" the change of the network's prediction score to pixels in the input image, leading to a high-resolution attribution map.

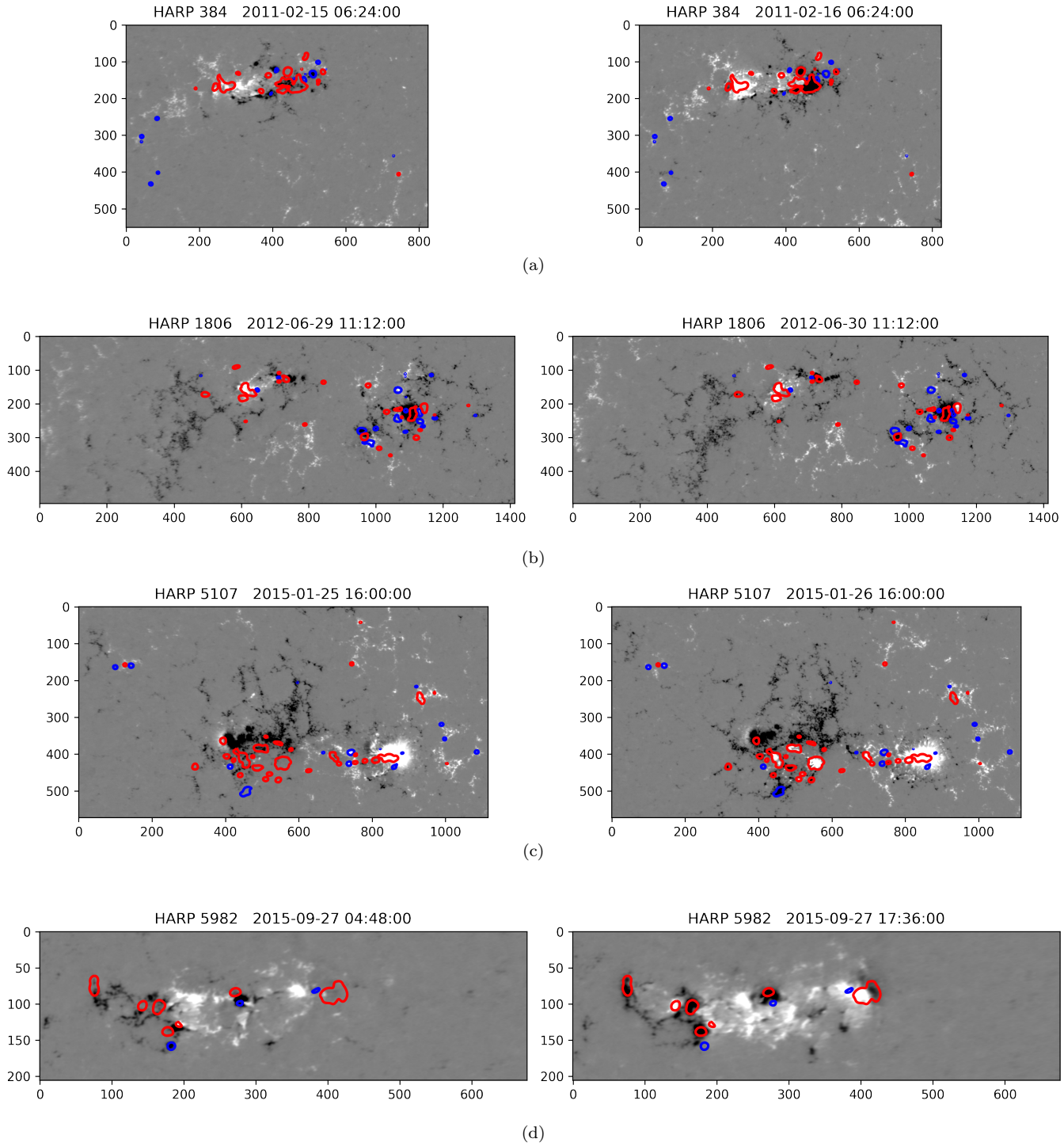
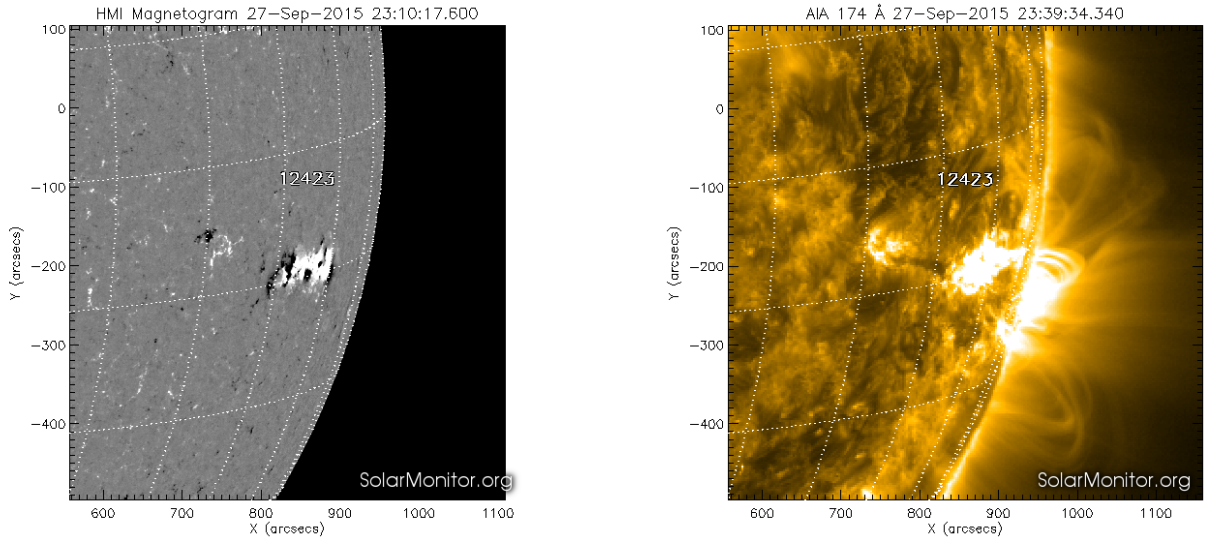


Figure 15. Highly attributed pixels in the last frame by Integrated Gradients on four select HARPs shown in rows. In (a), the left/right panel shows the first/last magnetogram in the sample sequence of HARP 384. The magnetograms are in the SHARP resolution, with ticks on the axes indicating pixels. Pixel values saturate at ±500 Gs. The red/blue contours on the right panel (last frame) highlight the areas with strong positive/negative Integrated Gradients relative to the first frame. The same contours are mapped to the left panel (first frame) for contrast. The contours are drawn on the attribution map smoothed with a Gaussian kernel with a standard deviation of 3 pixels. Figures in (b), (c), and (d) are similar to (a) but for other HARPs. The movies showing the evolution of Integrated Gradients of the entire sample sequence can be accessed at, e.g., https://zeyusun.github.io/attribution/contours/5107/contour_movie.gif for HARP 5107.



(a) HMI magnetogram

(b) AIA 171 Å image

Figure 16. Line-of-sight magnetic field (a) and solar EUV image (b) of HARP 5982 (NOAA AR 12423) at 23:10:17 on Sep 27, 2015. Images are taken from https://solarmonitor.org/. Note that the image title of (b) should be "AIA 171 Å" instead of "AIA 174 Å".

5. CONCLUSIONS AND DISCUSSION

In this paper, we used two solar cycles of active region observational data from SMARP (Bobra et al. 2021) and SHARP to examine the improvement in flare predictive performance of two deep learning models, namely the LSTM and the CNN, when trained on the fused datasets. When tested on SMARP, both models showed significant improvement. When tested on SHARP, LSTM showed significant improvement. The results of the controlled comparative studies indicate such an improvement is due to the significantly increased sample size from the other solar cycle. Then, in our setting of flare prediction, we verified the performance of the LSTM and the CNN using skill scores, reliability diagrams, ROC, and skill score profiles. The comparison showed that the LSTM is generally a better model than the CNN. After that, we explored the possibility of combining the LSTM and the CNN for a better prediction performance in the framework of a meta-learning paradigm called stacking. The results showed that in some settings, the stacking model outperforms the best member in the ensemble. Lastly, we applied visual attribution methods to CNNs. The results demonstrate the utility of visual attribution methods in identifying flare-related signatures in active regions, including the flux emergence and new polarity inversion lines. The attribution map on one particular region on the limb of the solar disk revealed one limitation of the CNN and suggested potential modifications for improvement.

The questions raised in Section 4 are arguably broad and general. We have taken one particular path to partially address each question. To inspire future studies, we provide additional comments and discussions related to these questions.

Task-based sample selection—In this work, we studied the task of distinguishing M- and X-class flare producing regions from flare-quiet regions. This "strong-vs-quiet" task focuses on the increase or the continuation of flare activity and does not require distinguishing between flares of closer energy levels like M- and C-class flares. The samples that indicate a decay in flare activity and the samples that only lead to weak flares are therefore excluded. This ensures good baseline classification performance against which our proposed predictors could be reliably compared. Our findings may not extend to other task definitions, e.g., all-clear forecasts, where flare-quiet active regions that evolve from a flare-active state are of interest (Barnes et al. 2016; Ji et al. 2020); and operationally-evaluated forecasts, where flare activities in the prediction period are considered as prescience and can not be used to select samples. The principal challenge to extending our analysis to these tasks is that weak- and no-flare activity annotations have higher uncertainty due to background radiation and other factors (McCloskey et al. 2018). The higher levels of "label noise" when including weak flares as negative samples would make learning a reliable predictor substantially more difficult.

A possible solution, and topic of future work, is to predict the continuous flare intensity level instead of the GOES flare activity class.

Evaluation under a realistic event rate—As mentioned in Section 2.5, we rebalance the training and the validation set to prevent the predictor biasing towards the majority non-event class simply due to its volume, and we rebalance the test set to evaluate the predictor's generalization ability under the same climatological rate as it is trained. Evaluating the performance under a realistic event rate requires more work other than simply applying the predictor on a test set that is not rebalanced: predictors trained on the balanced dataset will bias towards the minority class on a test set under a realistic rare event rate, producing an undesirably high false alarm rate. One possible solution to correct such bias is to treat the class proportions as priors and apply the Bayes rule (Elkan 2001). This method requires an accurate estimate of the true event rate of the testing period.

The importance and challenges of data fusion—Fusing data from multiple sources to produce more consistent, accurate, and useful information is a universal problem in astronomy. Although the astrophysics community is funding projects like DKIST (Rimmele et al. 2020) and the Vera Rubin Observatory (Ivezić et al. 2019), both of which will take 25–50 TB of data a day, astronomers cannot study long-term trends without including historical or old data sets (or waiting a decade for these instruments to take enough data). In this work, we took a straightforward approach to add the new data in the training set with minimal calibration, and train the models as usual. Based on our experiments, we have shown that this simple approach can result in improvement. We anticipate that, with more accurate cross-calibration between the SMARP and SHARP, the benefit of combining them may be better than demonstrated here. There are several possible ways to improve upon the fusion method:

- To simulate the effect of unresolved structures in SMARP magnetograms, Gaussian blur can be applied to the higher quality SHARP magnetogram. This approach is used in comparing full-disk line-of-sight magnetograms of HMI and MDI in Liu et al. (2012), in which the parameters of the Gaussian filters are tuned to minimize the root mean squared difference between them.

- Point spread functions can be estimated for MDI and HMI magnetograms separately, and deconvolution can be performed to remove stray light that is instrument-specific (Mathew et al. 2007; Yeo et al. 2014).

- Magnetogram fusion can be performed in the other direction: super-resolving magnetograms in SMARP to mimic those in SHARP. Such an approach has been recently explored using deep neural networks (Gitiaux et al. 2019; Jungbluth et al. 2019). The improved overall image quality of super-resolved SMARP magnetograms could capture higher resolution magnetic field distributions and hence improve the accuracy of the active region summary parameters in SMARP.

- For the active region summary parameters, we took a "post-facto" correction approach by correcting the parameters of the same name in the two data products via linear regression. Alternatively, with fused magnetograms available, one can also re-compute the parameters on those transformed image data. This approach avoids the linear assumption and leads to parameters more consistent with the manipulated magnetograms, with the caveat that the manipulated magnetograms also suffer from the loss of information. More concretely, the effects of spatial resolution on the inferred magnetic field and derived quantities have been examined by Leka & Barnes (2012), who found that, to preserve the underlying character of the magnetic field, post-facto binning can be employed with some confidence, albeit less so for derived quantities like vertical current density. In short, a universal and accurate fusing strategy that accounts for the instrumental spatial resolution is still hindered by our ignorance of the ground truth magnetic field structure, and the benefits and drawbacks of different fusing methods have to be evaluated case by case.

Machine learning with multi-source data—Learning from multi-source data is also a prevalent topic in machine learning. In our work, machine learning models are trained as usual with new data added to the training set. An alternative approach would use transfer learning: train on the additional data first, then switch to the original data for fine-tuning. In heliophysics, this idea is recently explored by Covas (2020) in the prediction of the solar surface longitudinally averaged radial magnetic field distribution, using historical data from 1874 to 1975 in addition to newer data obtained by SoHO and SDO.

Performance comparison between the LSTM and the CNN—The active region summary parameters used by the LSTM are derived from magnetograms. In that sense, the data used by the CNN contains complete information of the data used by the LSTM. However, our experiments show that the LSTM generally has better performance. There are many potential reasons that the CNN does not perform better than, or as well as the LSTM: (1) the CNN takes in uniformly sized magnetograms whose size and aspect ratio are distorted. (2) the CNN only uses the image of the last frame in the sequence, whereas the LSTM uses all the data in the sequence; (3) the CNN learns the features by itself, wheres the LSTM uses hand-crafted parameterizations that are known to be relevant to flaring activity; (4) the CNN uses subsampled images with information loss, whereas the LSTM uses parameters derived from full resolution images; (5) the CNN has more parameters and more prone to overfitting (which reflects on the lower training loss but not validation loss of the CNN in many experiments).

On the comparison of flare forecast methods—Many flare forecast studies quote the skill scores directly from other studies for comparison. Even though the forecast goal is somewhat standard and used in many studies (e.g. to predict whether there will be an M1.0+ class flare occurring in the future 24 hours), to conclude the superiority of one method against another, both methods have to be evaluated on the same dataset. However, it is not trivial to come up with such a "common ground" for methods to compete because research codes are not usually publicly available, and because different opinions exist on the ways the data should be processed. The difficulty in methodical comparison spawns the effort in fairly comparing existing forecasts (e.g., the "All-Clear" workshops (Barnes et al. 2016; Leka et al. 2019)) and developing common datasets (e.g., SWAN-SF benchmark dataset (Angryk et al. 2020)) or platforms (e.g., the FLARECAST project (Georgoulis et al. 2021)). To advocate credible comparisons, we do not quote skill scores from other studies. Instead, we follow the above studies and make our code publicly available to facilitate future comparisons.

On stacking ensemble—In our experiments, stacking CNN and LSTM performs similarly to the "select best" strategy but not significantly better in most settings. However, another stacking study Guerra et al. (2020) used a larger number of base learners and showed that most ensembles achieved a better skill score (between 5% to 15%) than any of the members alone. This suggests that improved performance may be obtained by training a larger number of base learners on the SMARP/SHARP data studied here.

Choice of the baseline in interpretability methods—Some visual attribution methods require reference input, such as Integrated Gradients and DeepLIFT. One naive choice is an image with all values equal to zero. Images of this sort imply a lack of patterns. These are the baselines mostly used for interpretation in computer vision tasks like object detection. In our case, the images are magnetic field component measurements, which can take on positive or negative values and a wide dynamic range, unlike normal images in real life. We choose the first image in the sequence as the reference, so that the visual attribution methods can attribute the change of prediction scores to the change of magnetic field configuration, which is of actual interest. There are other choices of baselines. One example is input images with Gaussian noise. Using this type of reference may reveal the sensitivity of the network's prediction to local changes. Furthermore, integration may benefit from going beyond simply linearly interpolating the reference and the input on the original image space, i.e., the 2D cartesian plane. For example, one could consider applying attribution methods to the path of time series of magnetograms. The Integrated Gradients calculated with this approach would integrate temporal dependency of each point-in-time in the sequence, exploiting more information about the evolution of active regions.

The authors would like to thank K. D. Leka for valuable discussions on the polarity artifacts of the line-of-sight component of the photospheric magnetic field, and on the effect of spatial resolution on magnetograms and derived quantities. This work was supported by NASA DRIVE Science Center grant 80NSSC20K0600.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., et al. 2018, in Advances in Neural Information Processing Systems, ed. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett, Vol. 31 (Curran Associates, Inc.). https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

Ahmadzadeh, A., Aydin, B., Georgoulis, M. K., et al. 2021, The Astrophysical Journal Supplement Series, 254, 23, doi: 10.3847/1538-4365/abec88

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. 2013, Int. J. Advance Soft Compu. Appl, 5

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. 2018, in International Conference on Learning Representations

Angryk, R. A., Martens, P. C., Aydin, B., et al. 2020, Scientific data, 7, 1

Barnes, G., Leka, K., Schrijver, C., et al. 2016, The Astrophysical Journal, 829, 89

Bhattacharjee, S., Alshehhi, R., Dhuri, D. B., & Hanasoge, S. M. 2020, The Astrophysical Journal, 898, 98

Bickel, P. J., & Doksum, K. A. 2015, Mathematical statistics: basic ideas and selected topics, volumes I-II package (CRC Press)

Bloomfield, D. S., Higgins, P. A., McAteer, R. J., & Gallagher, P. T. 2012, The Astrophysical Journal Letters, 747, L41

Bobra, M. G., & Couvidat, S. 2015, The Astrophysical Journal, 798, 135

Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, Solar Physics, 289, 3549

Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. 2021, The Astrophysical Journal Supplement Series, 256, 26, doi: 10.3847/1538-4365/ac1f1d

Bottou, L., Curtis, F. E., & Nocedal, J. 2018, Siam Review, 60, 223

Breiman, L. 1996, Machine learning, 24, 49

Campi, C., Benvenuto, F., Massone, A. M., et al. 2019, The Astrophysical Journal, 883, 150

Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, Space Weather, 17, 1404

Cinto, T., Gradvohl, A. L. S., Coelho, G. P., & da Silva, A. E. A. 2020, Monthly Notices of the Royal Astronomical Society, 495, 3332

Cohen, J. 1960, Educational and psychological measurement, 20, 37

Covas, E. 2020, Astronomische Nachrichten, 341, 384

Deng, Z., Wang, F., Deng, H., et al. 2021, The Astrophysical Journal, 922, 232, doi: 10.3847/1538-4357/ac2b2b

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018, arXiv preprint arXiv:1810.04805

Dua, D., & Graff, C. 2017, UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Džeroski, S., & Ženko, B. 2004, Machine learning, 54, 255

Elkan, C. 2001, in In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 973–978

Florios, K., Kontogiannis, I., Park, S.-H., et al. 2018, Solar Physics, 293, 1

Georgoulis, M. K., Bloomfield, D. S., Piana, M., et al. 2021, Journal of Space Weather and Space Climate, 11, 39, doi: 10.1051/swsc/2021023

Gitiaux, X., Maloney, S. A., Jungbluth, A., et al. 2019, arXiv preprint arXiv:1911.01486

Guerra, J. A., Murray, S. A., Bloomfield, D. S., & Gallagher, P. T. 2020, Journal of Space Weather and Space Climate, 10, 38

Guerra, J. A., Pulkkinen, A., & Uritsky, V. M. 2015, Space Weather, 13, 626

Hada-Muranushi, Y., Muranushi, T., Asai, A., et al. 2016, arXiv preprint arXiv:1606.01587

He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778

Hochreiter, S., & Schmidhuber, J. 1997, Neural computation, 9, 1735

Huang, X., Wang, H., Xu, L., et al. 2018, The Astrophysical Journal, 856, 7

Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, doi: 10.3847/1538-4357/ab042c

Ji, A., Aydin, B., Georgoulis, M. K., & Angryk, R. 2020, in 2020 IEEE International Conference on Big Data (Big Data), 4218–4225, doi: 10.1109/BigData50022.2020.9377906

Johnson, J. M., & Khoshgoftaar, T. M. 2019, Journal of Big Data, 6, 1

Jolliffe, I. T., & Stephenson, D. B. 2012, Forecast verification: a practitioner's guide in atmospheric science (John Wiley & Sons)

Jonas, E., Bobra, M., Shankar, V., Hoeksema, J. T., & Recht, B. 2018, Solar Physics, 293, 1

Jungbluth, A., Gitiaux, X., Maloney, S. A., et al. 2019, arXiv preprint arXiv:1911.01490

Kingma, D. P., & Ba, J. 2014, arXiv preprint arXiv:1412.6980

Krawczyk, B. 2016, Progress in Artificial Intelligence, 5, 221

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, Advances in neural information processing systems, 25, 1097

Kubo, Y. 2019, Journal of Space Weather and Space Climate, 9, A17

LeBlanc, M., & Tibshirani, R. 1996, Journal of the American Statistical Association, 91, 1641

Leka, K., & Barnes, G. 2003, The Astrophysical Journal, 595, 1296

—. 2012, Solar Physics, 277, 89

Leka, K., Barnes, G., & Wagner, E. 2017, Solar Physics, 292, 36

Leka, K., Park, S.-H., Kusano, K., et al. 2019, The Astrophysical Journal Supplement Series, 243, 36

Li, X., Zheng, Y., Wang, X., & Wang, L. 2020, The Astrophysical Journal, 891, 10

Liu, C., Deng, N., Wang, J. T., & Wang, H. 2017, The Astrophysical Journal, 843, 104
Liu, H., Liu, C., Wang, J. T., & Wang, H. 2019, The Astrophysical Journal, 877, 121
Liu, Y., Hoeksema, J., Scherrer, P., et al. 2012, Solar Physics, 279, 295
Mathew, S., Pillet, V. M., Solanki, S., & Krivova, N. 2007, Astronomy & Astrophysics, 465, 291
McCloskey, A. E., Gallagher, P. T., & Bloomfield, D. S. 2018, Journal of Space Weather and Space Climate, 8, A34
Murphy, A. H. 1973, Journal of Applied Meteorology and Climatology, 12, 595
Murray, S. A. 2018, Space Weather, 16, 777
Nie, W., Zhang, Y., & Patel, A. 2018, in International Conference on Machine Learning, PMLR, 3809–3818
Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., & Ishii, M. 2020, The Astrophysical Journal, 899, 150
Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. 2018, The Astrophysical Journal, 858, 113
Nishizuka, N., Sugiura, K., Kubo, Y., et al. 2017, The Astrophysical Journal, 835, 156
Nocedal, J., & Wright, S. 2006, Numerical optimization (Springer Science & Business Media)
Ribeiro, F., & Gradvohl, A. 2021, Astronomy and Computing, 35, 100468
Riley, P., Ben-Nun, M., Linker, J., et al. 2014, Solar Physics, 289, 769
Rimmele, T. R., Warner, M., Keil, S. L., et al. 2020, Solar Physics, 295, 1
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, SoPh, 162, 129, doi: 10.1007/BF00733429
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, SoPh, 275, 229, doi: 10.1007/s11207-011-9842-2
Schrijver, C. J. 2007, The Astrophysical Journal, 655, L117, doi: 10.1086/511857
Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in Proceedings of the IEEE international conference on computer vision, 618–626
ŞEn, M. U., & Erdogan, H. 2013, Pattern Recognition Letters, 34, 265
Shrikumar, A., Greenside, P., & Kundaje, A. 2017, in International Conference on Machine Learning, PMLR, 3145–3153
Silver, D., Huang, A., Maddison, C. J., et al. 2016, nature, 529, 484
Simonyan, K., Vedaldi, A., & Zisserman, A. 2013, arXiv preprint arXiv:1312.6034
Simonyan, K., & Zisserman, A. 2014, arXiv preprint arXiv:1409.1556
Springenberg, J., Dosovitskiy, A., Brox, T., & Riedmiller, M. 2015, in ICLR (workshop track)
Steward, G., Lobzin, V., Cairns, I. H., Li, B., & Neudegg, D. 2017, Space Weather, 15, 1151
Sundararajan, M., Taly, A., & Yan, Q. 2017, in Proceedings of the 34th International Conference on Machine Learning-Volume 70, 3319–3328
The SunPy Community, Barnes, W. T., Bobra, M. G., et al. 2020, The Astrophysical Journal, 890, 68, doi: 10.3847/1538-4357/ab4f7a
Ting, K. M., & Witten, I. H. 1999, Journal of artificial intelligence research, 10, 271
Todorovski, L., & Džeroski, S. 2003, Machine learning, 50, 223
Wang, X., Chen, Y., Toth, G., et al. 2020, The Astrophysical Journal, 895, 3
Wilks, D. S. 2011, Statistical methods in the atmospheric sciences, Vol. 100 (Academic press)
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. 2016, Data Mining: Practical Machine Learning Tools and Techniques (Morgan Kaufmann)
Wolpert, D. H. 1992, Neural networks, 5, 241
Woodcock, F. 1976, Monthly Weather Review, 104, 1209
Xue, J.-H., & Hall, P. 2014, IEEE transactions on pattern analysis and machine intelligence, 37, 1109
Yeates, A. R. 2020, Solar physics, 295, 1
Yeo, K., Feller, A., Solanki, S., et al. 2014, Astronomy & Astrophysics, 561, A22
Yi, K., Moon, Y.-J., Lim, D., Park, E., & Lee, H. 2021, The Astrophysical Journal, 910, 8
Yu, D., Huang, X., Wang, H., et al. 2010, The Astrophysical Journal, 710, 869
Yuan, Y., Shih, F. Y., Jing, J., & Wang, H.-M. 2010, Research in Astronomy and Astrophysics, 10, 785
Zeiler, M. D., & Fergus, R. 2014, in European conference on computer vision, Springer, 818–833
Zheng, Y., Li, X., & Wang, X. 2019, The Astrophysical Journal, 885, 73