Embedding Ordinality to Binary Loss Function for Improving Solar Flare Forecasting

Chetraj Pandey, Anli Ji, Jinsu Hong, Rafal A. Angryk, Berkay Aydin
Dept. of Computer Science, Georgia State University, Atlanta, GA, USA
{*cpandey1, aji1, jhong36, rangryk, baydin2*}*@gsu.edu*

2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA) | 979-8-3503-6494-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/DSAA61799.2024.10722839

***Abstract*—Several natural phenomena, such as floods, earthquakes, volcanic eruptions, or extreme space weather events often come with severity indexes. While these indexes, whether linear or logarithmic are vital, data-driven predictive models for these events rather use a fixed threshold. In this paper, we explore encoding this ordinality to enhance the performance of data-driven models, with specific application in solar flare forecasting. The prediction of solar flares is commonly approached as a binary forecasting problem, categorizing events as either Flare (FL) or No-Flare (NF) based on a chosen threshold (e.g., ≥C-class, ≥M-class, or ≥X-class). However, this binary formulation overlooks the inherent ordinality between the sub-classes within each binary class (FL and NF). In this paper, we propose a novel loss function aimed at optimizing the binary flare prediction problem by embedding the intrinsic ordinal flare characteristics into the binary cross-entropy (BCE) loss function. This modification is intended to provide the model with better guidance based on the ordinal characteristics of the data and improve the overall performance of the models. For our experiments, we employ a ResNet34-based model with transfer learning to predict ≥M-class flares by utilizing the shape-based features of magnetograms of active region (AR) patches spanning from -90° to $+90^{\circ}$ of solar longitude as our input data. We use a composite skill score (CSS) as our evaluation metric, which is calculated as the geometric mean of the True Skill Score (TSS) and the Heidke Skill Score (HSS) to rank and compare our models' performance. The primary contributions of this work are as follows: (i) We introduce a novel approach to encode ordinality into a binary loss function showing an application to solar flare prediction, (ii) We enhance solar flare forecasting by enabling flare predictions for each AR across the entire solar disk, without any longitudinal restrictions, and evaluate and compare performance. (iii) Our candidate model, optimized with the proposed loss function, shows an improvement of ∼7%, ∼4%, and ∼3% for AR patches within $\pm 30^{\circ}$, $\pm 60^{\circ}$, and $\pm 90^{\circ}$ of solar longitude, respectively in terms of CSS, when compared with standard BCE. Additionally, we demonstrate the ability to issue flare forecasts for ARs in near-limb regions (regions between $\pm 60^{\circ}$ to $\pm 90^{\circ}$) with a CSS=0.34 (TSS=0.50 and HSS=0.23), expanding the scope of AR-based models for solar flare prediction. This advances the reliability of solar flare forecasts, leading to more effective prediction capabilities.**

***Index Terms*—Solar flares, Deep learning, Optimization**

I. Introduction

From earthquakes to tornadoes, and volcanic eruptions to extreme space weather events, natural occurrences that pose hazards to our society often come with a severity index. This index may follow a linear scale (such as flood severity [1] or tornadoes [2]) or a logarithmic one (for example, earthquakes [3], volcanic activity [4], or space weather events like flares or solar energetic particle events [5]). Predictive models for these events commonly incorporate a set threshold; however, these models can gain advantages from incorporating the ordinality of these severity indices. In this work, we will delve into encoding this information appropriately to efficiently optimize data-driven predictive models, with specific applications to binary solar flare forecasting.

Solar flares are short-lived events on the Sun observed as intense outbursts of energy radiating from the Sun's surface in the form of extreme ultraviolet and X-ray radiation, and they are the central phenomena in space weather forecasting. They are classified according to their peak X-ray flux level into the following five categories by the National Oceanic and Atmospheric Administration (NOAA): X ($> 10^{-4} Wm^{-2}$), M ($> 10^{-5} Wm^{-2}$), C ($> 10^{-6} Wm^{-2}$), B ($> 10^{-7} Wm^{-2}$), and A ($> 10^{-8} Wm^{-2}$) [5]. These five major flare classes are measured on a logarithmic scale and ordered as X>M>C>B>A. Flares weaker than A-class are generally undetectable and are classified as flare-quiet (FQ). M-class and X-class solar flares are rare events and much more powerful than other flare classes. These stronger flares (M- and X-class) attract the attention of researchers because they can potentially impact conditions near Earth and disrupt technological systems such as satellite communications, GPS navigation, power grids, and aviation [6]. Therefore, solar flare prediction in a binary setting is most commonly formulated as predicting ≥M-class flares.

Ordinality-aware Loss Function: In solar flare forecasting, the binary prediction framework involves categorizing flares based on their flare magnitude. Specifically, setting the threshold at ≥M categorizes M- and X-class flares as Flare (FL), while FQ-, A-, B-, and C-class flares are designated as No Strong Flare (NF). This approach simplifies prediction by distinguishing significant flares (M- and X-class) from less intense activity, aiding in assessing potential solar disturbances. However, the intrinsic ordinal flare characteristics in sub-class level is overlooked during model optimization. Traditional loss functions like cross-entropy and focal loss [7], commonly used in data-driven model optimization for binary settings, cannot account for ordinal characteristics within the FL and NF classes. They treat all instances equally during optimization, failing to distinguish between different sub-classes within these categories. This approach does not fully utilize the ordinal information inherent in the flare classification system. Hence, in our

work, we propose encoding these ordinal characteristics as weighting factors in the binary cross-entropy (BCE) loss function. By doing so, we assign different weights to instances based on their flare sub-class, ensuring that the model optimizes for the specific nuances within each class. We hypothesize that this adjustment is particularly relevant considering the ordinal nature of the flare events classification.

Flare Forecasting with Projection Effects: In this study, we utilize images of line-of-sight (LoS) magnetograms of active region (AR) patches. ARs are high activity regions on the Sun's surface, distinguished by their intense magnetic fields concentrated within sunspots. These magnetic fields often undergo significant distortion and instability, triggering plasma disturbances and releasing energy in the form of flares and other solar phenomena [8]). This makes ARs the regions of interest emphasizing the importance of utilizing AR-based features for predicting solar flares, as the disturbed magnetic fields in them are often linked as the main initiators of these solar events. However, the magnetic field measurements, which are the main features used in AR-based forecasting techniques, are susceptible to severe projection effects caused by the orientation of the observing instrument relative to the solar surface. Therefore, as ARs get closer to limbs to the degree that after $\pm 60^{\circ}$ of solar longitude, the magnetic field readings are distorted [9], which limits the existing models to include data pertaining to central locations only [10]. To address this, we derive images from original LoS magnetogram rasters of AR patches.

Data Preprocessing Pipeline for ARs: Our data preprocessing pipeline introduced in [11], converts high-dimensional magnetic field rasters to images capturing the overall morphology and spatial distribution of ARs retaining the important shape-based parameters such as size, directionality, sunspot borders, and polarity inversion lines [12]. Shape-based features retained in these derived images of magnetograms provide a robust representation of the overall underlying magnetic activity. We recognize the persistence of severe projection effects, however, we hypothesize that the complex feature learning capabilities of contemporary deep learning models can potentially learn from these distorted readings. Consequently, we include data encompassing ARs in near-limb regions (beyond $\pm 60^{\circ}$) as well, thereby offering a novel capability to predict solar flares throughout the entire solar disk.

Furthermore, it is essential to note that the tracked AR patches vary in size depending on the size of the ARs. Existing approaches have been limited to AR patches in central locations, often resizing rectangular patches to obtain square images. However, this resizing distorts the original aspect ratio, consequently altering the shapes and sizes of ARs. Alternatively, variable-sized AR patches are cropped (using methods like center crop or random crop) to obtain square images, resulting in information loss. In contrast, we proposed and utilized a sliding window kernel-based approach in [11]. This method select such a cropped region that maximizes total unsigned flux (USFLUX: the sum of the absolute of the magnetic field strength values), maintaining the original aspect ratios of AR patches and preserving critical spatial features. By maximizing the USFLUX, we ensure that we extract the most representative region with significant magnetic flux build up. This method adapts to the variability in AR patch shapes and sizes, avoiding distortion and prioritizing the capture of more relevant information.

Leveraging the images of LoS magnetograms of AR patches, we develop a predictive model for solar flares of magnitude $\geq$M-class. We employ these images to train ResNet34 [13] based model with different configurations of our proposed loss function. Our contributions can be summarized as follows: (i) We introduce a novel approach to encode inherent ordinality of data to binary cross-entropy (BCE) loss function, showing the effectiveness in solar flare prediction as a case study, (ii) We show that our models are capable to predict flares across the entire solar disk, including often overlooked near-limb regions, improving the comprehensiveness of AR-based solar flare prediction models. This study presents a key advancement in the field of solar flare prediction, contributing to ongoing efforts aimed at enhancing space weather forecasting capabilities and improving our understanding of solar phenomena.

The remainder of the paper is structured as follows. Sec. II provides an overview of existing studies on solar flare predictions using deep learning models and various data sources. In Sec. III, we detail the process of data collection with labeling and consequent data distribution, and describe the architecture of our flare prediction model. In Sec. IV we outline our methodology by providing a detailed description of our modification to the standard cross-entropy loss and its application to solar flare prediction. Sec. V presents the experimental design and hyperparameter configurations of our model. Sec. VI presents our model evaluation showing the effectiveness of our approach on models' performance evaluated with skill scores. Finally, in Sec. VII, we summarize our findings and suggest avenues for future research.

II. Related Work

Several approaches such as human-based predictions (e.g., [14]), statistical models (e.g., [15]), and numerical simulations based on physics-based models (e.g., [16]), have been employed to predict solar flares. Recently, the use of data-driven approaches, which leverage machine learning and deep learning techniques, has significantly increased (e.g., [17]–[20]) owing to their capacity to exploit extensive datasets and their experimental achievements in space weather forecasting [21]. As solar flares are phenomena caused by sudden, abrupt changes in the magnetic field in the solar atmosphere, these data-driven approaches most commonly utilize magnetogram-based data which includes solar full-disk magnetograms (e.g., [10], [22]–[25]), multivariate time series (MVTS) data extracted from solar vector magnetograms (e.g., [26], [27], [28]), cutouts or patches of tracked AR (e.g., [29], [30]), and features summarizing each AR patch (e.g., [31], [32]).

A deep learning model based on a multi-layer perceptron to predict solar flares $\geq$C and $\geq$M class was presented in [33]. In this study, they used manually selected features extracted from multi-modal solar observations of the full solar disk, which included vector magnetograms and extreme ultraviolet (EUV) images to predict $\geq$M- and $\geq$C-class flares. In [34], a convolutional neural network (CNN) based hybrid model is proposed to predict the occurrence of $\geq$C-class flares. Similarly, [24], [35] presented a CNN-based model to predict $\geq$M-class flares utilizing full-disk magnetogram images. While these full-disk models include near-limb regions, by design they are unable to localize the relevant ARs that are likely to flare and instead issue one single forecast for the entire solar disk.

In [32], a support vector machine based model trained with 25 AR summary parameters extracted from vector magnetograms of AR patches within $\pm 68^{\circ}$ of solar longitude was presented. Similarly, in [27], a deep learning based time series classifier and in [28] a sliding window Time Series Forest (TSF) was trained with a MVTS data of 24 space weather related physical parameters primarily calculated from AR magnetograms within $\pm 70^{\circ}$ of solar longitude. Furthermore, a CNN-based flare forecasting model trained with solar AR patches (resized to 100×100 pixels) extracted from LoS magnetograms within $\pm 30^{\circ}$ of solar longitude to predict $\geq$C-, $\geq$M-, and $\geq$X-class flares was presented in [29]. More recently, [36] proposed a CNN-based model named "CARFFM-4" trained with AR patches (sized to 160×160 pixels) created from R parameter [37] within $\pm 30^{\circ}$ of solar longitude to predict $\geq$M-class flares in the next 48 hours.

All the literature reviewed in this section, formulates solar flare prediction as a binary forecasting problem utilizing a binary loss function without any flare ordinal characteristics. Furthermore, It is important to note that, there is variability in the literature in terms of the type of data modality which includes multiple instruments (HMI/SDO, AIA/SDO, MDI/SOHO) and data types (EUV images, magnetograms and extracted features corresponding to AR and full-disk). Furthermore the variability in prediction targets ($\geq$C-, $\geq$M-, $\geq$X-class flares) and forecasting horizon (24 hours and 48 hours) is also prominent. The predictive capabilities of AR-based models are often limited by observations taken from central locations from $\pm 30^{\circ}$ to $\mp 70^{\circ}$. The full-disk models complement the issue of longitudinal coverage in AR-based models; however, they fail to pin-point an active region and issue a single forecast for the entire solar disk. In this work, we introduce a new loss function to build a limb-to-limb flare prediction model that is trained on magnetogram images of AR-patches spanning full 180° ($\pm 90^{\circ}$) of solar longitude and evaluate our models' efficacy in different zones defined by longitudinal range and provide a novel capability, to our knowledge, missing in operational systems.

III. Data and Model

The primary raw input data in our work are obtained from line-of-sight (LOS) magnetograms of ARs provided by the Helioseismic and Magnetic Imager (HMI) [38] onboard the Solar Dynamics Observatory (SDO) [39], which are publicly available as a data product named Spaceweather HMI Active Region Patches (SHARP) [40] from the Joint Science Operations Center[1] at a temporal cadence of 12 minutes. In this work, we utilized magnetograms spanning from May 2010 to December 2018, sampling magnetograms at a cadence of one hour. The magnetograms of AR patches contain rasters of magnetic field strength values typically ranging from $\sim\pm 4500$ G. An example of magnetogram of an AR patch is shown in Fig. 1 (a). Along with magnetograms, we use bitmaps (another data product from the SHARP series) which define the region with pixels located within or outside the ARs, providing the region of interest within the AR patch as shown in Fig. 1 (b). The bitmaps are equal in size to the LOS magnetograms of AR patches, where the area represented by white pixels shows the region within the AR and hence our region of interest [40]. For each AR patch, we assign a binary label using peak X-ray flux converted to NOAA/GOES flare classes such that: (i) $\geq$M indicates Flare (FL) signifying the existence of a relatively strong flaring activity, and (ii) <M indicates No Flare (NF) with a prediction window of 24 hours. To illustrate, from the timestamp of an AR patch to the next 24 hours, if the maximum NOAA/GOES flare class is <M, then we label the AR patch as NF; otherwise, FL.

Fig. 1. An illustrative example of (a) Original raw input magnetogram of HMI AR patch corresponding to HARP number: 7115 (NOAA AR number: 12673) observed on 2017-09-06 at 06:00:00 UTC, (b) Bitmap corresponding to HMI AR patch in (a) showing the high activity region (region of interest) indicated by white pixels, (c) Final processed image of AR patch in (a) now sized to 512×512, that is used to train our models.

In our data processing pipeline, introduced in our prior work [11] and illustrated in Fig. 2, we begin by collecting hourly instances of raw input magnetograms of AR patches, alongside their corresponding bitmaps. Our initial step involves applying the bitmap as a filter to precisely crop the AR patches, isolating the regions with high activity. Subsequently, we implement a size filter: if the resulting cropped AR patches are smaller than 70 pixels in width, we exclude them from our dataset. It is worth noting that we determine this threshold based on the overall data distribution, ensuring retention of all instances corresponding to 'FL' instances while removing those from the 'NF' class. Following this filtering stage, we proceed to adjust the magnetic flux. We cap the flux values at $\pm$256G, and any flux values within $\pm$25G are set to 0 to mitigate noise. Ensuring uniformity in size, we apply zero-padding to patches smaller than 512×512 pixels. Conversely, for larger patches, exceeding 512×512 pixels, we employ a 512×512 kernel to

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

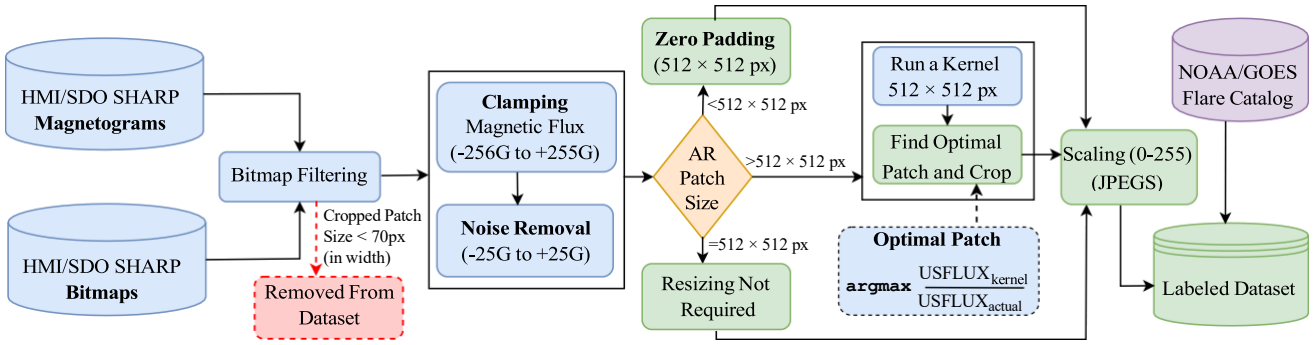


Fig. 2. The overall schema of the data processing pipeline used in this work. It shows a sequential pipeline for creating JPEG images from magnetogram rasters and corresponding bitmaps used for cropping the regions with relevant information. Boxes colored in green collectively defines our dataset.

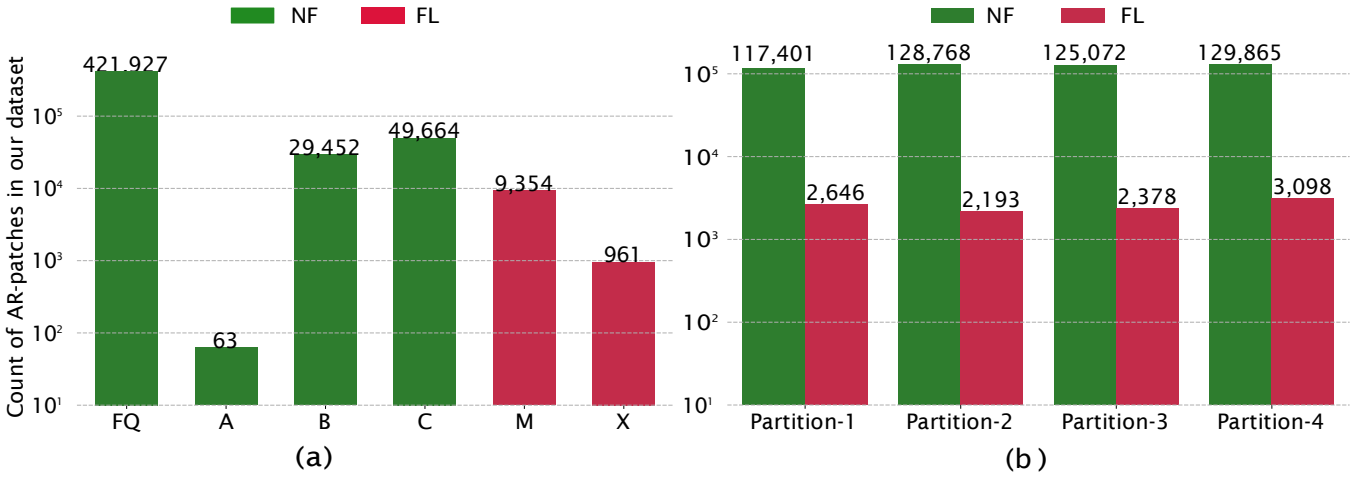
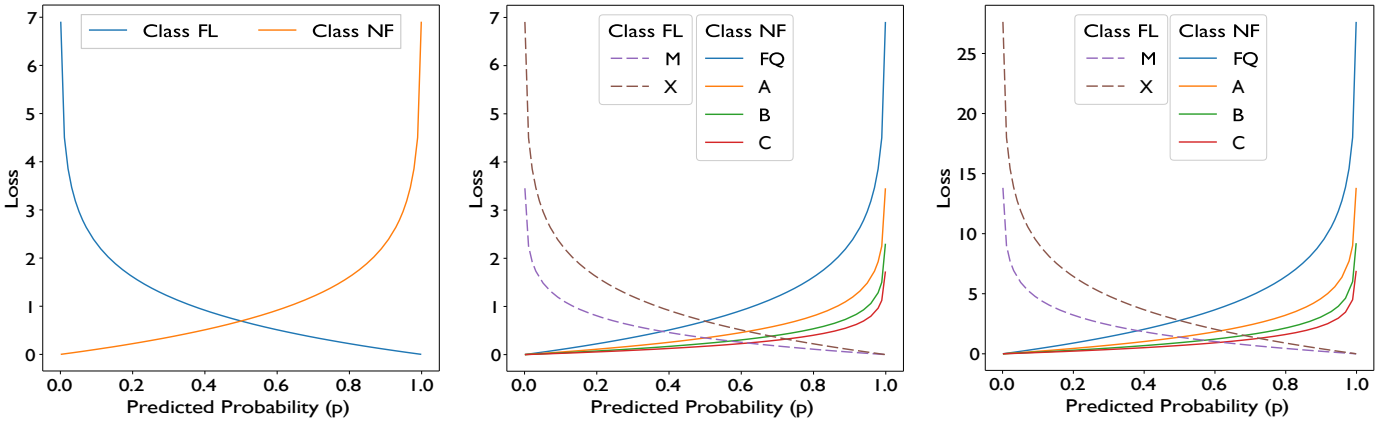


Fig. 3. (a) A bar plot representing the overall distribution of flare classes in our dataset (b) A bar plot showing binarized (≥M) flare distributions across the four temporally non-overlapping tri-monthly data partitions. Note: The height of the bars are in logarithmic scale.

select the patch with the maximum total unsigned flux (namely USFLUX, which is the sum of the absolute value of magnetic field strength represented as raster values in magnetograms). By doing this, we aim to minimize information loss by picking a spatial window where the total flux is the highest, which is more likely to include the regions of interest. Finally, to standardize the representation, all patches are scaled to fit within the range of 0-255, facilitating the generation of images. An example of a final processed image is shown in Fig. 1 (c) generated using the magnetogram of AR patch in Fig. 1 (a) and the corresponding bitmap in Fig. 1 (b), providing an illustration of the outcomes of our data preprocessing steps.

The overall distribution of our labeled AR patches data, with binary flare classes NF (comprising flare-quiet (FQ), A-, B-, and C-class flares) and FL (including M- and X-class flares), is shown in Fig. 3 (a). In total, we have 501,106 instances belonging to the NF class and 10,315 instances belonging to the FL class, resulting in a class imbalance ratio of $\sim$ 1:49. We split our dataset into four non-overlapping tri-monthly partitions as shown in Fig. 3 (b), using the onset timestamps of the HARP series to ensure that each AR trajectory remains entirely within a single partition, thus avoiding any overlap. This approach contrasts with the method described in [41], which uses magnetogram observation timestamps for partitioning the full-disk magnetograms. Finally, we use Partitions 1 and 2 as our training set while Partitions 3 and 4 are used as validation and test set respectively. The preprocessed dataset used in this study is publicly available from [42].

The task of solar flare prediction in this work is formalized as a binary image classification problem; therefore, we select a general CNN based model, ResNet34 [13]. Recently, attention-based models, notably Vision Transformers (ViTs) [43], have gained prominence for their superior performance in image classification tasks. Despite their state-of-the-art results, these models tend to have a high number of trainable parameters, making them resource-heavy and less suitable for applications with limited computational resources or small datasets. For our specific application involving a small dataset, we opted for a more straightforward approach using a CNN-based model, specifically ResNet34. We modified the ResNet34 architecture to handle 1-channel input magnetogram images (grayscale images) by adding an initial convolutional layer with a 3×3 kernel, a stride of 1, and three output feature maps. This modification allows the model to utilize pre-trained weights effectively while processing the 1-channel magnetogram images. The final architecture includes 34 convolutional layers



(a) BCE LOSS (b) BCE-SF LOSS (α=1) (c) BCE-SF LOSS (α=4)

Fig. 4. An illustrative plot showing: (a) Standard binary cross-entropy (BCE) loss. (b-c) BCE for solar flare prediction (BCE-SF) which encodes ordinal flare characteristics as loss weighting mechanism with α=1 and α=4 respectively. Note: FL class indicates target 1 and NF class indicates target 0.

(including all residual and basic convolutional layers), one max pooling layer, one adaptive average pooling layer, and one fully connected layer.

IV. Ordinality-aware Loss Function Design

In this work, we utilize a novel loss function designed for binary solar flare prediction that encodes the ordinal flare characteristics in the standard binary cross-entropy (BCE) loss. Let N be the total number of instances in a batch. Let y_i denote the true label for the i-th sample, where $y_i \in \{0, 1\}$. Let p_i be the predicted probability that the i-th sample belongs to the "FL" class (target 1), defined as $p_i = \sigma(\hat{y}_i)$, where $\hat{y}_i$ is the model output (logit) and σ is the sigmoid function, then the standard binary cross-entropy BCE$(y, \hat{y})$ loss function is represented as shown in Eq. (1) and the corresponding loss function plot is shown in Fig. 4 (a).

$$\text{BCE}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \tag{1}$$

As mentioned earlier, in the binary setting of solar flare prediction where our chosen threshold is ≥M, the two binary classes are: (i) the NF-class including FQ-, A-, B-, and C-class instances, and (ii) the FL-class including M- and X-class instances. As each of these individual flare classes are ordinal in nature, where FQ<A<B<C<M<X, we introduce a weighting factor based on the corresponding sub-classes (flare classes within each binary class) such that instances belonging to these sub-classes are represented by weights β as shown in Eq. (2).

$$\beta = \begin{cases} 10 & \text{if FQ} \\ 10^2 & \text{if A} \\ 10^3 & \text{if B} \\ 10^4 & \text{if C} \\ 10^2 & \text{if M} \\ 10 & \text{if X} \end{cases} \tag{2}$$

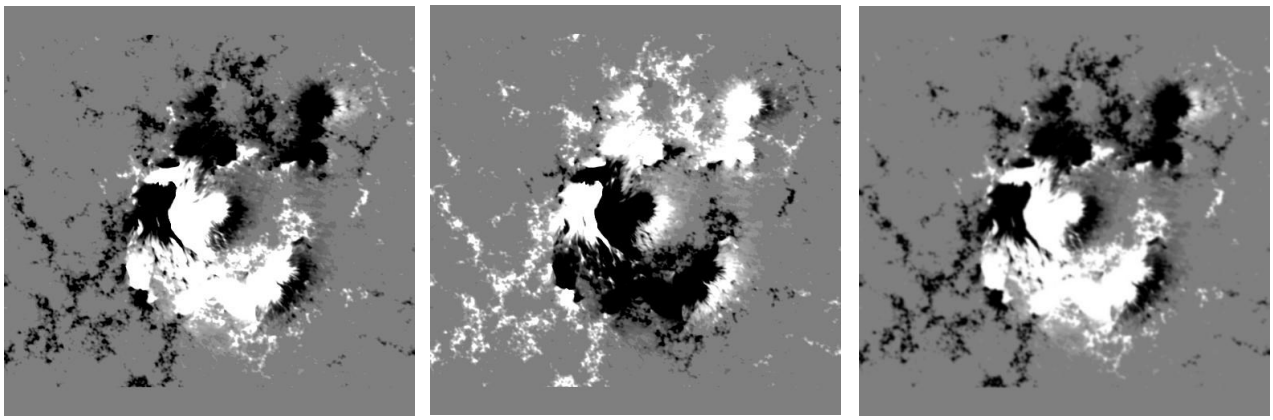
The BCE-SF loss is designed in such a way that the incorrect predictions in the binary flare classes result in different losses based on the sub-class of the instances. To elaborate further, since an X-class flare is ten times more powerful than an M-class flare, even though both belong to class FL, the loss value for an X-class flare should be higher than that for an M-class flare. Similarly, a B-class flare in the NF class is ten times weaker than C-class flares in the same binary class. Therefore, incorrect predictions of these two classes should have different loss values. Thus, we used these ordinal weights (β_i) representing individual flare classes, and our proposed binary cross-entropy loss for solar flare prediction (BCE-SF) can be represented as shown in Eq. (3).

$$\text{BCE-SF}(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} a \times \text{BCE}(y_i, \hat{y}_i) \times \frac{1}{\log_{10}(\beta_i)} \tag{3}$$

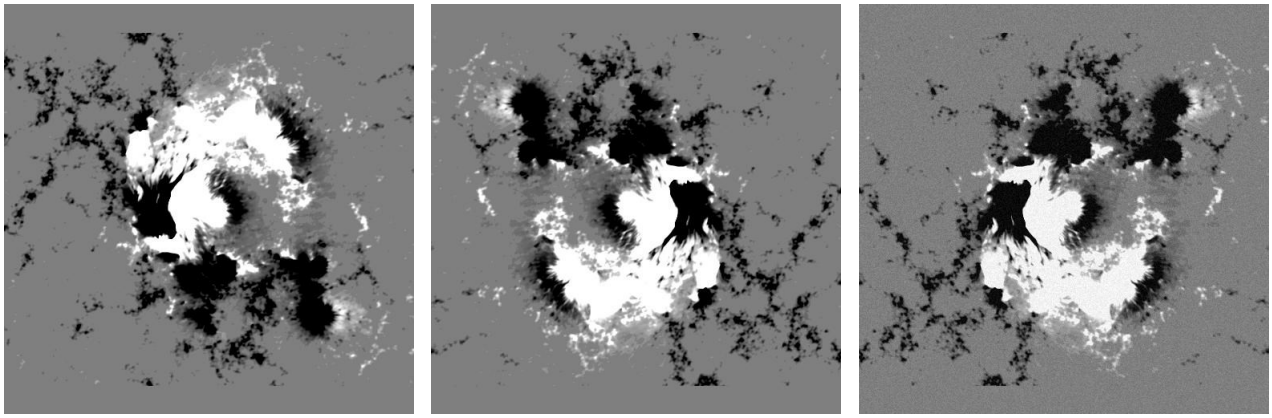
Here, a is a scaling factor that aligns the loss values with the scale of the corresponding BCE loss. Specifically, when $a = 1$, the maximum loss value for an incorrectly predicted instance matches the scale of the BCE loss, as shown in Fig. 4 (b). In this case, the loss value scale for FQ- and X-class (the two extremes of the binary categories) aligns with the corresponding BCE loss, while all other incorrect predictions have lower loss values. Similarly, when $a = 4$, the minimum loss value for an incorrectly predicted instance matches the scale of the BCE loss, as shown in Fig. 4 (c). Here, the loss value scale for C-class aligns with the BCE loss scale, while all other incorrect predictions have higher loss values. Therefore, we recommend the range of $a \in [1, 4]$ which can be regarded as a hyperparameter for optimal performance. It is important to note that the BCE-SF loss, by leveraging intrinsic flare characteristics, offers a simple modification to the BCE loss without introducing new model-dependent parameters.

V. Experimental Settings

In this section, we comprehensively delve into our dataset preparation methods for model training and evaluation, along-



(a) Original (Preprocessed) (b) Polarity Inversion (c) Gaussian Blurring



(d) Vertical Flipping (e) Horizontal Flipping (f) Add Noise (+25G)

Fig. 5. An illustrative example of (a) input magnetogram of HMI AR patch corresponding to HARP number: 7115 (NOAA AR number: 12673). (b-f) five different augmentations applied to AR patch in (a). These augmentations are applied to the processed magnetograms before scaling to 0-255.

side detailing our model configurations in regard to the usage of BCE and BCE-SF loss functions, and hyperparameters. Furthemore, we provide the definition of our evaluation metrics and the rationale behind the selection of these metrics.

A. Dataset

As mentioned earlier in Sec. III, we follow time-segmented trimonthly partitioning to create four partitions of our entire dataset. Partition-1 and 2 combined are used as the training set. However, due to significant class imbalance in our dataset, we used undersampling together with data augmentation to create a balanced training set. Firstly, we augmented data instances belonging to the FL-class in our training set using five data augmentation techniques: (i) polarity inversion, which swaps the signs of positive polarity to negative and vice versa as shown in Fig. 5 (b), (ii) Gaussian filtering, which applies a Gaussian blur to the image to reduce noise and detail (Fig. 5 (c)), (iii) Vertical Flipping , which involves flipping the image along a horizontal axis (Fig. 5 (e)), (iv) Horizontal Flipping, which involves flipping the image along a vertical axis (Fig. 5 (d)), and (v) Adding random noise within$\pm$25G (Fig. 5 (f)). To balance the FL-Class instances with NF, we undersampled our training data by randomly selecting 30% of instances belonging to A-, B-, C-class flares each, and ~8% of instances from FQ from both Partition-1 and 2. For realistic evaluation, we maintained the original imbalanced distribution in Partitions 3 and 4, which are our validation and test sets respectively, as shown in Fig. 6.

B. Model Parameters

TABLE I
HYPERPARAMETERS SEARCH SPACE WITH EXPERIMENTALLY OBSERVED OPTIMAL HYPERPARAMETERS FOR EACH MODEL.

Hyperparameters	Search Space	Optimal Parameters	
		BCE	BCE-SF
Initial Learning Rate	{0.00001 to 0.01}	0.01	0.001
Weight Decay	{0.00001 to 0.01}	0.01	0.001
Batch Size	{48, 64, 80}	64	64
Scaling Factor (α)	{1, 2, 3, 4}	N/A	2

In our model hyperparameter selection process, we define the hyperparameter space, encompassing initial learning rates (η), weight decay parameters, batch sizes, and scaling factors (α) as shown in Table I. Following the definition of our hyperparameter space, we conduct a grid search across this space, evaluating on the validation set for all our models.

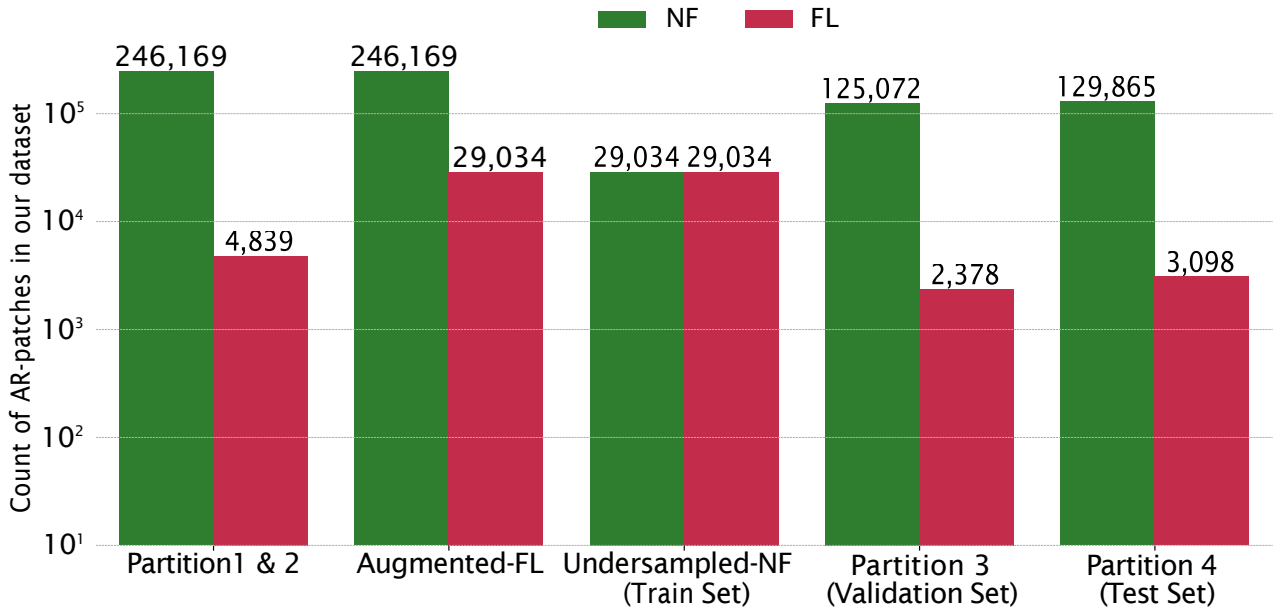


Fig. 6. The overall distribution of data instances partitioned into train set (showing original, augmented and undersampled data counts), validation set, and test set used in this work.

During this search, we train our models using stochastic gradient descent (SGD) with BCE and BCE-SF loss function. Additionally, we employ a dynamic learning rate strategy called `ReduceLROnPlateau` with a factor of 0.3 and a patience period of 2 epochs. This learning rate scheduling mechanism starts the training with an initial learning rate (η) as mentioned in Table. I. If the validation loss does not improve for two consecutive epochs (patience period), the new learning rate is calculated as follows:

$$\eta_{\text{new}} := \eta_{\text{current}} \times \text{factor}$$

Upon completing the grid search and evaluating the models, we identified the optimal hyperparameters as shown in Table I. These parameters exhibited superior performance during the search and we use these to train our final models for 50 epochs and evaluate on the test set.

C. Evaluation Metrics

True Skill Statistic (TSS, in Eq. 4) and Heidke Skill Score (HSS, in Eq. 5), derived from the four elements of confusion matrix: TP, TN, FP, FN are the two forecast skills scores widely used in evaluating flare prediction models.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \tag{4}$$

$$HSS = 2 \times \frac{TP \times TN - FN \times FP}{((P \times (FN + TN) + (TP + FP) \times N))} \tag{5}$$

where, $N = TN + FP$ and $P = TP + FN$.

The values of TSS and HSS range from -1 to 1, where 1 indicates all correct predictions, -1 represents all incorrect predictions (also, it means that all inverse predictions are correct, i.e., there is a skill), and 0 represents no skill. Unlike TSS, HSS is a metric that accounts for class imbalance. It is commonly used in evaluating solar flare prediction models because these datasets typically have a high imbalance ratio as discussed in [44], [45]. However, choosing a candidate model based on two skill scores becomes difficult, as it demands preference of one metric over another at the end. Therefore, by combining TSS and HSS in a geometric mean as in the Composite Skill Score (CSS, in Eq. 6), we obtain a single metric that balances between discrimination ability and imbalance awareness.

$$CSS = \begin{cases} 0, & \text{if } TSS \times HSS < 0 \\ \sqrt{TSS \times HSS}, & \text{otherwise} \end{cases} \tag{6}$$

CSS considers both the discrimination power of the model (TSS) and its ability to outperform random chance (HSS), offering a more comprehensive evaluation. It provides a single metric that accounts for both aspects of model performance, making it more suitable for assessing forecast models, particularly in scenarios with class imbalance. Hence, we evaluate and compare our models based on the single metric, which is CSS but report both TSS and HSS for completeness. For reproducibility, the source codes for this work is publicly available from our open source repository [46].

VI. Experimental Evaluation

As explained earlier in Sec. V-A, we conducted experiments to predict solar flares in a binary setting ($\geq$M-class flares) using a "train-validation-test split" of our entire dataset, which consists of magnetograms of AR patches covering a solar longitudinal range of $\pm 90^\circ$ (i.e., the entire solar disk). We utilized the validation set to monitor the models' performance every epoch and tuned hyperparameters to optimize the CSS.

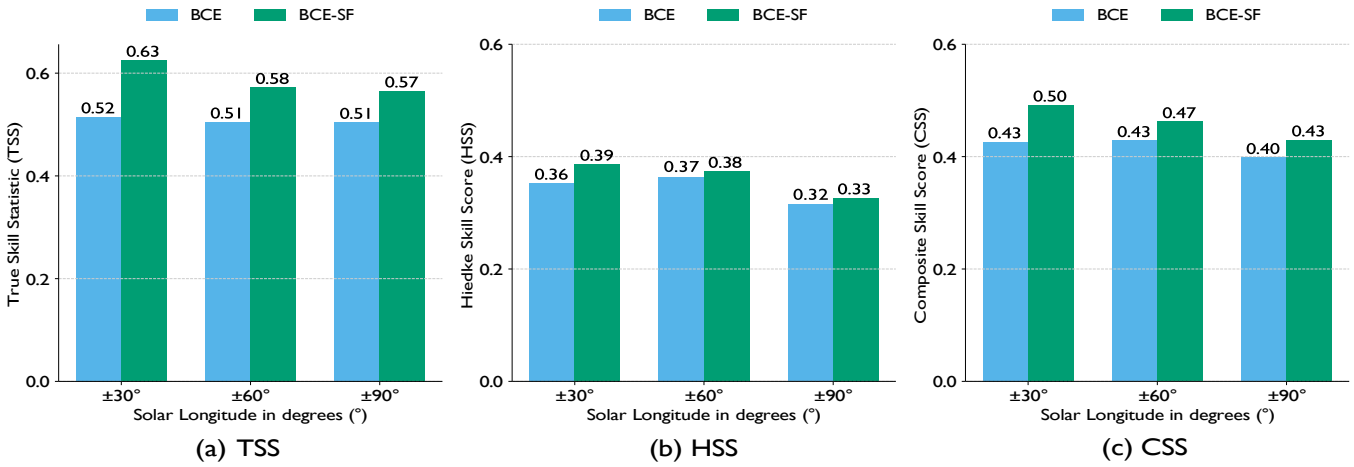


Fig. 7. Performance of our models trained with both BCE and BCE-SF loss was evaluated on the **test set** in terms of (a) TSS, (b) HSS, and (c) CSS. The skill scores are shown for different ranges of solar longitudes: ±30°, ±60°, and ±90°. This longitudinal ranges indicates the performance evaluated on ARs from the test set located within 0° to 30°, 0° to 60°, and 0° to 90° in both the directions (East (-ve) and West (+ve)) of the Sun.

After training the model with optimal hyperparameters, we employed a threshold tuning approach to calibrate our models by tuning the prediction score thresholds. This involved evaluating the performance of each model on the validation set at different threshold values ranging from 0.01 to 0.99 with an increment of 0.01. We selected the threshold that optimized CSS for each model individually. However, this resulted into a same threshold value of 0.69 for both the models trained with BCE and BCE-SF loss functions. These thresholds were then applied to the test set for both models, and the models' performance was reported on the test set.

Additionally, we assessed the performance of both of our models on subsets of data representing different longitudinal coverages: within $\pm 30^{\circ}$, $\pm 60^{\circ}$, and $\pm 90^{\circ}$ of solar longitude. To elaborate, longitudinal coverage of $\pm 30^{\circ}$ indicates AR patches in our test set that correspond to the central region of the Sun, encompassing up to -30° in the East direction and up to $+30^{\circ}$ in the West direction from the center (0°). A coverage of $\pm 60^{\circ}$ extends further, encompassing a broader region that extends 60° to the East and West of the solar longitude. This range covers a significant portion of the Sun's surface, allowing for a more comprehensive examination of solar activity beyond just the central region but still within a reasonably close proximity to it. A coverage of $\pm 90^{\circ}$ encompasses the entire test set. The performance of our models relative to each other in terms of TSS, HSS, and CSS is illustrated in Fig. 7 (a), (b), and (c), respectively.

Upon evaluation, we noted that the model trained with the BCE-SF loss function consistently outperformed the one optimized with BCE loss across all three skill scores and longitudinal coverages. Comparing the performance on the longitudinal coverage of $\pm 30^{\circ}$, we observed $\sim$11%, $\sim$3%, and $\sim$7% higher skill scores in terms of TSS, HSS, and CSS, respectively, as shown in Fig. 7. Furthermore, this improvement was also observed with increased longitudinal coverage. For ARs in the test set within $\pm 60^{\circ}$, we noted $\sim$7%, $\sim$1%, and $\sim$4% higher TSS, HSS, and CSS, respectively, with BCE-SF compared to the model trained with BCE. Overall, for our entire test set, we observed $\sim$ 6%, $\sim$ 1%, and $\sim$3% higher scores in terms of TSS, HSS, and CSS, respectively, for the BCE-SF trained model over BCE, as illustrated in Fig. 7, indicated by $\pm 90^{\circ}$. Additionally, our analysis revealed a linearly decreasing trend in model performance with increasing longitudinal coverage of ARs, with highest skill scores noted for ARs within ($\pm 30^{\circ}$) and lowest when within $\pm 90^{\circ}$) for both of the models (BCE and BCE-SF).

After noticing the pattern indicating a decline in model performance with increasing longitudinal coverage, we investigated the effectiveness of our models on non-overlapping regions of solar longitudes. To facilitate this analysis, we delineated three zones: (i) within 0° to $\pm 30^{\circ}$, (ii) the region between $\pm 30^{\circ}$ to $\pm 60^{\circ}$, and (iii) the region between $\pm 60^{\circ}$ to $\pm 90^{\circ}$. Similar to our earlier evaluation across overlapping longitudinal ranges, we computed all three skill scores to evaluate the model's performance across these zones. In doing so, we observed a similar linearly decreasing trend in skill scores as our earlier evaluation, highest in central regions (0° to $\pm 30^{\circ}$) and lowest in limb regions ($\pm 60^{\circ}$ to $\pm 90^{\circ}$), as illustrated in Fig. 8 from all three models. Interestingly, we observed that while TSS is 3% lower, HSS was 1% higher with the BCE-trained model compared to BCE-SF when evaluating within $\pm 30^{\circ}$ to $\pm 60^{\circ}$ of the solar longitude, as shown in Fig. 8 (b). This highlights using a composite skill score, as choosing the model based solely on TSS or HSS scores might lead to a false sense of good performance.

It is worth noting that while existing models are typically designed to predict solar flares up to $\pm 60^{\circ}$ of the solar longitude, our model demonstrates capability in the near-limb regions, namely the region between $\pm 60^{\circ}$ to $\pm 90^{\circ}$. Despite having lower skill scores compared to those in the central region, this study reveals a new capability that demonstrates skill in the near-limb region, thereby advancing solar flare prediction. This advancement underscores the significance of our research in extending predictive models for solar flares

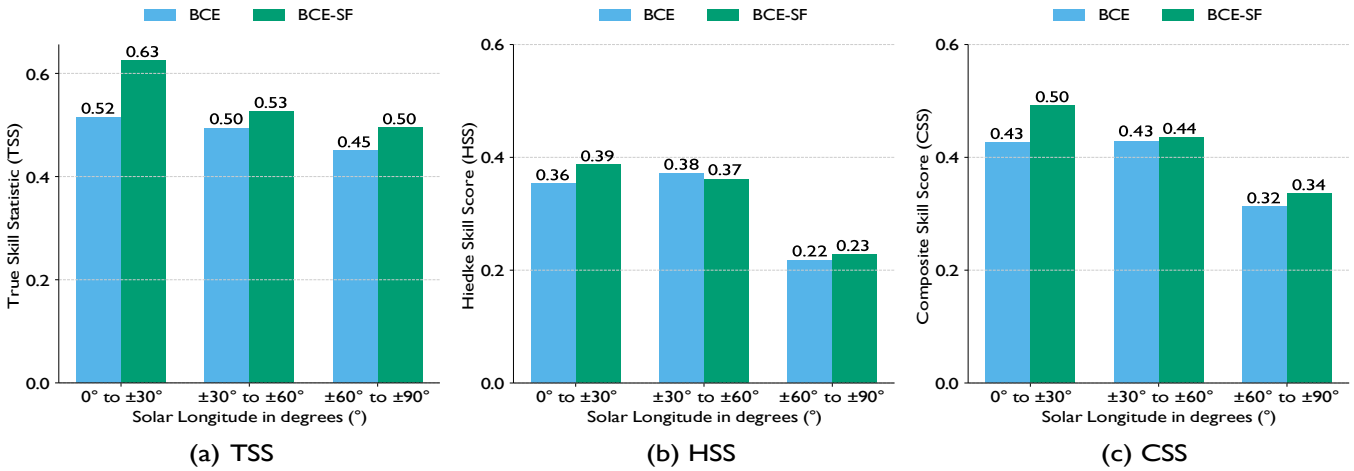


Fig. 8. Performance of our models trained with both BCE and BCE-SF loss was evaluated on the **test set** in terms of (a) TSS, (b) HSS, and (c) CSS. The skill scores are shown for different ranges of solar longitudes with **non-overlapping** regions: 0° to ±30°, ±30° to ±60°, and ±60° to ±90°. This longitudinal ranges indicates the performance evaluated on ARs from the test set located within 0° to 30°, 30° to 60°, and 60° to 90° in both the directions (East (-ve) and West (+ve)) of the Sun.

beyond central regions, thereby improving our understanding and forecasting capabilities for solar phenomena.

VII. Conclusion and Future Work

In this study, we primarily introduced a novel ordinality-aware binary loss function to optimize data-driven predictive models and demonstrated its effectiveness in improving predictive capabilities compared to the standard binary loss function, specifically in the application of solar flare prediction. Furthermore, using our data preprocessing pipeline, we utilized AR patches encompassing the limb-to-limb range of the Sun (i.e., $\pm$ 90°) to build flare prediction models capable of forecasting solar flares of magnitude $\geq$ M-class. Upon evaluating the capability of our limb-to-limb models, the results show that we can satisfactorily predict flaring activity despite severe projection effects, although there is room for improvement (the skill is limited when compared to central locations). Additionally, the results show that shape-based features in magnetograms are effective for predicting solar flares even when the ARs are close to the limbs.

While full-disk models are developed to complement AR-based models in near-limb regions, they lack the ability to localize AR-specific predictions. We define this work as an important step towards fully integrating ARs into solar flare prediction, with implications for advancing such predictions. Numerous avenues for future exploration exist, which include exploration on utilizing the actual peak X-ray fluxes as instance ordinality into the binary loss functions, investigating this approach with multi-modal solar observations, developing spatiotemporal models, and incorporating explanatory and interpretative frameworks into the model to enhance reliability.

Acknowledgments

This work is supported in part under two NSF grants (Award #2104004 and #1931555) and a NASA SWR2O2R grant (Award #80NSSC22K0272). The data used in this study is a courtesy of NASA/SDO and the AIA, EVE, and HMI science teams, and the NOAA National Geophysical Data Center (NGDC).

References

[1] A. J. Schroeder, J. J. Gourley, J. Hardy, J. J. Henderson, P. Parhi, V. Rahmani, K. A. Reed, R. S. Schumacher, B. K. Smith, and M. J. Taraldsen, "The development of a flash flood severity index," *Journal of Hydrology*, vol. 541, p. 523–532, Oct. 2016. [Online]. Available: http://dx.doi.org/10.1016/j.jhydrol.2016.04.005

[2] C. A. Doswell, H. E. Brooks, and N. Dotzek, "On the implementation of the enhanced fujita scale in the usa," *Atmospheric Research*, vol. 93, no. 1–3, p. 554–563, Jul. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.atmosres.2008.11.003

[3] D. M. Boore, "The richter scale: its development and use for determining earthquake source parameters," *Tectonophysics*, vol. 166, no. 1–3, p. 1–14, Sep. 1989. [Online]. Available: http://dx.doi.org/10.1016/0040-1951(89)90200-X

[4] C. G. Newhall and S. Self, "The volcanic explosivity index (vei) an estimate of explosive magnitude for historical volcanism," *Journal of Geophysical Research: Oceans*, vol. 87, no. C2, p. 1231–1238, Feb. 1982. [Online]. Available: http://dx.doi.org/10.1029/JC087iC02p01231

[5] L. Fletcher, B. R. Dennis, H. S. Hudson, S. Krucker, K. Phillips, A. Veronig, M. Battaglia, L. Bone, A. Caspi, Q. Chen, P. Gallagher, P. T. Grigis, H. Ji, W. Liu, R. O. Milligan, and M. Temmer, "An observational overview of solar flares," *Space Science Reviews*, vol. 159, no. 1-4, pp. 19–106, Aug. 2011. [Online]. Available: https://doi.org/10.1007/s11214-010-9701-8

[6] Y. Yasyukevich, E. Astafyeva, A. Padokhin, V. Ivanova, S. Syrovatskii, and A. Podlesnyi, "The 6 september 2017 x-class solar flares and their impacts on the ionosphere, GNSS, and HF radio wave propagation," *Space Weather*, vol. 16, no. 8, pp. 1013–1027, Aug. 2018.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r, "Focal loss for dense object detection," 2017. [Online]. Available: https://arxiv.org/abs/1708.02002

[8] S. Toriumi and H. Wang, "Flare-productive active regions," *Living Reviews in Solar Physics*, vol. 16, no. 1, May 2019. [Online]. Available: https://doi.org/10.1007/s41116-019-0019-7

[9] D. A. Falconer, S. K. Tiwari, R. L. Moore, and I. Khazanov, "A new method to quantify and reduce the net projection error in whole-solar-active-region parameters measured from vector magnetograms," *The Astrophysical Journal*, vol. 833, no. 2, p. L31, Dec. 2016. [Online]. Available: https://doi.org/10.3847/2041-8213/833/2/l31

[10] C. Pandey, R. A. Angryk, and B. Aydin, *Explaining Full-Disk Deep Learning Model for Solar Flare Prediction Using Attribution Methods*. Springer Nature Switzerland, 2023, p. 72–89. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-43430-3_5

[11] C. Pandey, T. Adeyeha, J. Hong, R. A. Angryk, and B. Aydin, "Advancing solar flare prediction using deep learning with active region patches," 2024. [Online]. Available: https://arxiv.org/abs/2406.11054

[12] A. Ji, X. Cai, N. Khasayeva, M. K. Georgoulis, P. C. Martens, R. A. Angryk, and B. Aydin, "A systematic magnetic polarity inversion line data set from sdo/hmi magnetograms," *The Astrophysical Journal Supplement Series*, vol. 265, no. 1, p. 28, Mar. 2023. [Online]. Available: http://dx.doi.org/10.3847/1538-4365/acb43a

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[14] M. D. Crown, "Validation of the NOAA space weather prediction center's solar flare forecasting look-up table and forecaster-issued probabilities," *Space Weather*, vol. 10, no. 6, pp. n/a–n/a, Jun. 2012. [Online]. Available: https://doi.org/10.1029/2011sw000760

[15] K. Lee, Y.-J. Moon, J.-Y. Lee, K.-S. Lee, and H. Na, "Solar flare occurrence rate and probability in terms of the sunspot classification supplemented with sunspot area and its changes," *Solar Physics*, vol. 281, no. 2, pp. 639–650, Sep. 2012. [Online]. Available: https://doi.org/10.1007/s11207-012-0091-9

[16] K. Kusano, T. Iju, Y. Bamba, and S. Inoue, "A physics-based method that can predict imminent large solar flares," *Science*, vol. 369, no. 6503, pp. 587–591, Jul. 2020. [Online]. Available: https://doi.org/10.1126/science.aaz2511

[17] J. Hong, A. Ji, C. Pandey, and B. Aydin, *Beyond Traditional Flare Forecasting: A Data-driven Labeling Approach for High-fidelity Predictions*. Springer Nature Switzerland, 2023, p. 380–385.

[18] C. Pandey, R. A. Angryk, and B. Aydin, "Unveiling the potential of deep learning models for solar flare prediction in near-limb regions," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 703–708.

[19] J. Hong, A. Ji, C. Pandey, and B. Aydin, "Enhancing solar flare prediction with innovative data-driven labels," in *2023 IEEE 5th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, Nov. 2023.

[20] J. Hong, C. Pandey, A. Ji, and B. Aydin, "An innovative solar flare metadata collection for space weather analytics," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 408–413.

[21] K. Whitman, R. Egeland, and I. G. R. et al., "Review of solar energetic particle models," *Advances in Space Research*, Aug. 2022. [Online]. Available: https://doi.org/10.1016/j.asr.2022.08.006

[22] C. Pandey, A. Ji, R. A. Angryk, M. K. Georgoulis, and B. Aydin, "Towards coupling full-disk and active region-based flare prediction for operational space weather forecasting," *Frontiers in Astronomy and Space Sciences*, vol. 9, Aug. 2022.

[23] C. Pandey, R. A. Angryk, M. K. Georgoulis, and B. Aydin, *Explainable Deep Learning-Based Solar Flare Prediction with Post Hoc Attention for Operational Forecasting*. Springer Nature Switzerland, 2023, p. 567–581.

[24] C. Pandey, A. Ji, T. Nandakumar, R. A. Angryk, and B. Aydin, "Exploring deep learning for full-disk solar flare prediction with empirical insights from guided grad-cam explanations," in *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 2023, pp. 1–10.

[25] C. Pandey, A. Ji, R. A. Angryk, and B. Aydin, "Towards interpretable solar flare prediction with attention-based deep neural networks," in *2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2023, pp. 83–90.

[26] A. Ji, B. Aydin, M. K. Georgoulis, and R. Angryk, "All-clear flare prediction using interval-based time series classifiers," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2020, pp. 4218–4225. [Online]. Available: http://dx.doi.org/10.1109/BigData59044.2023.10386908

[27] A. Ji, J. Wen, R. Angryk, and B. Aydin, "Solar flare forecasting with deep learning-based time series classifiers," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, Aug. 2022.

[28] A. Ji and B. Aydin, "Interpretable solar flare prediction with sliding window multivariate time series forests," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, Dec. 2023.

[29] X. Huang, H. Wang, L. Xu, J. Liu, R. Li, and X. Dai, "Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms," *The Astrophysical Journal*, vol. 856, no. 1, p. 7, Mar. 2018. [Online]. Available: https://doi.org/10.3847/1538-4357/aaae00

[30] X. Li, Y. Zheng, X. Wang, and L. Wang, "Predicting solar flares using a novel deep convolutional neural network," *The Astrophysical Journal*, vol. 891, no. 1, p. 10, Feb. 2020. [Online]. Available: https://doi.org/10.3847/1538-4357/ab6d04

[31] D. S. Bloomfield, P. A. Higgins, R. T. J. McAteer, and P. T. Gallagher, "Toward reliable benchmarking of solar flare forecasting methods," *The Astrophysical Journal*, vol. 747, no. 2, p. L41, Feb. 2012.

[32] M. G. Bobra and S. Couvidat, "Solar flare prediction using≤SDO≤/HMI VECTOR MAGNETIC FIELD DATA WITH a MACHINE-LEARNING ALGORITHM," *The Astrophysical Journal*, vol. 798, no. 2, p. 135, Jan. 2015. [Online]. Available: https://doi.org/10.1088/0004-637x/798/2/135

[33] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, and M. Ishii, "Deep flare net (DeFN) model for solar flare prediction," *The Astrophysical Journal*, vol. 858, no. 2, p. 113, May 2018. [Online]. Available: https://doi.org/10.3847/1538-4357/aab9a7

[34] E. Park, Y.-J. Moon, S. Shin, K. Yi, D. Lim, H. Lee, and G. Shin, "Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms," *The Astrophysical Journal*, vol. 869, no. 2, p. 91, Dec. 2018.

[35] C. Pandey, R. A. Angryk, and B. Aydin, *Deep Neural Networks Based Solar Flare Prediction Using Compressed Full-disk Line-of-sight Magnetograms*. Springer International Publishing, 2022, p. 380–396.

[36] M. Li, Y. Cui, B. Luo, J. Wang, and X. Wang, "Deep neural networks of solar flare forecasting for complex active regions," *Frontiers in Astronomy and Space Sciences*, vol. 10, Jun. 2023.

[37] C. J. Schrijver, "A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting," *The Astrophysical Journal*, vol. 655, no. 2, p. L117–L120, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1086/511857

[38] J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat, M. C. Rabello-Soares, R. S. Bogart, J. T. Hoeksema, Y. Liu, T. L. Duvall, D. J. Akin, B. A. Allard, J. W. Miles, R. Rairden, R. A. Shine, T. D. Tarbell, A. M. Title, C. J. Wolfson, D. F. Elmore, A. A. Norton, and S. Tomczyk, "Design and ground calibration of the helioseismic and magnetic imager (HMI) instrument on the solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1-2, pp. 229–259, Oct. 2011. [Online]. Available: https://doi.org/10.1007/s11207-011-9842-2

[39] W. D. Pesnell, B. J. Thompson, and P. C. Chamberlin, "The solar dynamics observatory (SDO)," *Solar Physics*, vol. 275, no. 1-2, pp. 3–15, Oct. 2011. [Online]. Available: https://doi.org/10.1007/s11207-011-9841-3

[40] M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. D. Leka, "The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: Sharps – space-weather hmi active region patches," *Solar Physics*, vol. 289, no. 9, p. 3549–3578, Apr. 2014.

[41] C. Pandey, R. A. Angryk, and B. Aydin, "Solar flare forecasting with deep neural networks using compressed full-disk hmi magnetograms," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 1725–1730. [Online]. Available: https://doi.org/10.1109/bigdata52589.2021.9671322

[42] C. Pandey and B. Aydin, "Bitmap filtered line of sight hmi active region patches with augmentations," 2024. [Online]. Available: https://doi.org/10.7910/DVN/WLTBQQ

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[44] A. Ahmadzadeh, M. Hostetter, B. Aydin, M. K. Georgoulis, D. J. Kempton, S. S. Mahajan, and R. Angryk, "Challenges with extreme class-imbalance and temporal coherence: A study on solar flare data," in *2019 IEEE Intl. Conf. on Big Data (Big Data)*. IEEE, Dec. 2019.

[45] A. Ahmadzadeh, B. Aydin, M. K. Georgoulis, D. J. Kempton, S. S. Mahajan, and R. A. Angryk, "How to train your flare prediction model: Revisiting robust sampling of rare events," *The Astrophysical Journal Supplement Series*, vol. 254, no. 2, p. 23, May 2021. [Online]. Available: http://dx.doi.org/10.3847/1538-4365/abec88

[46] DMLab@GSU, "Ordinal loss sf," 2024, accessed: 2024-08-20. [Online]. Available: https://bitbucket.org/gsudmlab/ordinalloss-sf/src/main/