

Exploring Deep Learning for Full-disk Solar Flare Prediction with Empirical Insights from Guided Grad-CAM Explanations

Chetraj Pandey¹, Anli Ji¹, Trisha Nandakumar², Rafal A. Angryk¹, Berkay Aydin¹

^{1,2}*Dept. of Computer Science, Georgia State University, Atlanta, GA, USA*

¹*{cpandey1, aji1, rangryk, baydin2}@gsu.edu*

²*trishan104@gmail.com*

Abstract—This study progresses solar flare prediction research by presenting a full-disk deep-learning model to forecast $\geq$ M-class solar flares and evaluating its efficacy on both central (within $\pm 70^\circ$) and near-limb (beyond $\pm 70^\circ$) events, showcasing qualitative assessment of post hoc explanations for the model’s predictions, and providing empirical findings from human-centered quantitative assessments of these explanations. Our model is trained using hourly full-disk line-of-sight magnetogram images to predict $\geq$ M-class solar flares within the subsequent 24-hour prediction window. Additionally, we apply the Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) attribution method to interpret our model’s predictions and evaluate the explanations. Our analysis unveils that full-disk solar flare predictions correspond with active region characteristics. The following points represent the most important findings of our study: (1) Our deep learning models achieved an average true skill statistic (TSS) of ~ 0.51 and a Heidke skill score (HSS) of ~ 0.38 , exhibiting skill to predict solar flares where for central locations the average recall is ~ 0.75 (recall values for X- and M-class are 0.95 and 0.73 respectively) and for the near-limb flares the average recall is ~ 0.52 (recall values for X- and M-class are 0.74 and 0.50 respectively); (2) qualitative examination of the model’s explanations reveals that it discerns and leverages features linked to active regions in both central and near-limb locations within full-disk magnetograms to produce respective predictions. In essence, our models grasp the shape and texture-based properties of flaring active regions, even in proximity to limb areas—a novel and essential capability with considerable significance for operational forecasting systems.

Index Terms—Solar flares, Deep learning, xAI, Interpretability

I. INTRODUCTION

Solar flares are transient but intense outbursts of energy that emanate from the Sun’s surface in the form of extreme ultraviolet and X-ray radiation and are one of the central phenomena in space weather forecasting. Flares are typically classified in five different classes based on their peak X-ray flux by National Oceanic and Atmospheric Administration (NOAA) [1], where flare intensities are characterized by a major class (i.e., A, B, C, M, and X) and a numerical strength value within a class (from 1.0 to 9.9). These indicate the order of magnitude of the peak X-ray flux in a logarithmic scale (e.g., X-ray flux values for B3.7 is $3.7 \times 10^{-7} \text{ W m}^{-2}$, C7.2 is $7.2 \times 10^{-6} \text{ W m}^{-2}$, M1.4 is $1.4 \times 10^{-5} \text{ W m}^{-2}$, or X2.1 is $2.1 \times 10^{-4} \text{ W m}^{-2}$). M- and X-class solar flares are relatively scarce events and significantly more powerful than the other

flare classes which garner the attention of researchers due to their potential to cause near-Earth impacts that pose a substantial risk to both space-based and terrestrial infrastructures, disrupting satellite communications, power grids, and aviation, making solar flare prediction a complex and critical area of research in space weather forecasting [2].

Active regions (ARs) on the Sun represent areas where the sun’s magnetic field experiences disruptions, leading to a variety of solar phenomena such as solar flares, coronal mass ejections (CMEs), and solar energetic particle (SEP) events. The majority of solar flare prediction models concentrate on these regions of interest, offering forecasts for each AR, as they are primarily responsible for space weather events. To generate a comprehensive forecast using an AR-centric model, the likelihood of a flare occurring within each AR ($P_{FL}(AR_i)$) is usually consolidated utilizing a heuristic function as described in [3]. The equation for this aggregation is $P_{aggregated} = 1 - \prod_i [1 - P_{FL}(AR_i)]$, which computes the probability of having a minimum of one flaring AR, under the assumption that flares from different ARs occur independently and infers that all ARs hold equal significance in the combination. This constrains an accurate assessment of the full-disk flare prediction probability.

Furthermore, the magnetic field measurements, which underpin AR-based forecasting techniques, are influenced by projection effects when ARs are in proximity to the limbs [4]. Consequently, the data is restricted to ARs situated within a range of $\pm 30^\circ$ [5] to $\pm 70^\circ$ [6] of the solar disk due to significant projection effects [7]. The full-disk models, however, utilize magnetograms corresponding to the entire solar disk to rely on shape-based features such as size, directionality, shapes of sunspots, and polarity inversion lines similar to the findings from [8]–[11]. While projection effects persist in these images, we show that the full-disk models are capable of learning and predicting flare productivity of areas close to the limb of the Sun, offering an essential component for operational systems.

Recently, many studies (e.g., [3], [5], [12]–[18]) have led to the successful application of machine learning and deep learning methods in predicting solar flares. These methods have shown promising experimental results, indicating their potential to improve forecasting accuracy in the field of solar flare prediction. However, deep learning models are

often considered black-box models due to their complex data representations, making it challenging to understand the reasoning behind their predictions. This lack of transparency can pose problems in critical applications like solar flare prediction, where the model's reliability is crucial. The absence of transparency in the decision-making process lowers confidence in the accuracy and reliability of the predictions made. Post hoc explanations offer insights into a model's decision-making process, enabling the identification of potential errors and biases in the data, and subsequent improvements to the model's performance and reliability, which addresses the issue of transparency and makes it more trustworthy for critical applications. While post hoc explanations effectively highlight crucial input features and enhance trust, a gap persists in our understanding due to the inherent opacity of these models' decision-making processes. Therefore, an approach for meaningful translations remains essential for fully understanding these explanations. This paper serves as an initial step in shedding light on our models' effectiveness by clarifying the model's reasoning with post hoc explanations and enhancing confidence in the model's reliability and accuracy, especially in cases where incorrect predictions could result in substantial consequences, such as in space weather forecasting.

The focus of this paper is to study whether full-disk models for solar flare prediction can be relied upon for operational forecasting applications as this exploration extends to the prediction of near-limb flare events with model explanations, which hold a notably critical role. This paper builds upon the convolutional neural network (CNN) based model in [19] to predict $\geq M$ -class flares, trained with compressed full-disk line-of-sight magnetograms. The novel contributions of this paper are as follows: (i) We highlight the overall improved performance of our solar flare prediction model exhaustively with quantitative analysis of predictive performance on near-limb and central locations, (ii) We utilized a recent attribution method to explain and interpret the decisions of our deep learning model and present a case study that shows the explanations of model's decision in spatiotemporal progression, and (iii) More importantly, for the first time, we provide empirical evidence that our models can tackle the prediction of flares appearing on near-limb regions of the Sun by analyzing the explanations with quantitative human-centered approach [20].

The remainder of the paper is structured as follows: Section II provides an overview of existing studies on machine learning and deep learning-based data-driven solar flare prediction models. In Section III, we detail the process of data collection with labeling and explain the subsequent data distribution. In Section IV, we outline our methodology by describing the architecture of our flare prediction model and providing a detailed explanation of the method used for explanation. In Section V, we present our experimental design and findings from the model evaluation. Furthermore, we offer case-based qualitative interpretations of explanations for limited instances and a quantitative analysis of explanations for entire X-class flares in our dataset. Finally, in Section VI, we summarize our findings and suggest avenues for future research.

II. RELATED WORK

Currently, to the best of our knowledge, four strategies are employed for flare prediction: (i) empirical human prediction (e.g., [21], [22]), (ii) statistical prediction (e.g., [23], [24]), (iii) physics-based numerical simulations (e.g., [25], [26]), and (iv) machine learning and deep learning approaches (e.g., [3], [5], [14]–[18], [27]). With the rapid progress in machine learning and deep learning techniques, their application to solar flare prediction has significantly expedited research efforts. For instance, in [13], a multi-layer perceptron-based model was employed for predicting $\geq C$ - and $\geq M$ -class flares. The model utilized 79 manually selected physical precursors extracted from multi-modal solar observations.

Deep learning models have recently emerged as a prominent choice for solar flare prediction. In [5], a CNN-based model was trained using AR patches extracted from line-of-sight magnetograms within $\pm 30^\circ$ of the central meridian to predict $\geq C$ -, $\geq M$ -, and $\geq X$ -class flares. Similarly, [14] developed a CNN-based model that issued binary class predictions for $\geq C$ - and $\geq M$ -class flares within 24 hours using Space-Weather Helioseismic and Magnetic Imager Active Region Patches (SHARP) data [28]. The SHARP data was extracted from solar magnetograms using AR patches located within $\pm 45^\circ$ of the central meridian. Notably, both models had limited operational capability, as they were restricted to small portions of the observable disk in central locations ($\pm 30^\circ$ and $\pm 45^\circ$).

In the context of explainability in solar flare prediction models, [29] used an occlusion-based method to interpret a CNN-based solar flare prediction model trained with AR patches. Similarly, [30] presented visual explanation methods for a deep learning-based flare prediction model. They used Grad-CAM [31], and Guided Backpropagation [32] to show the relationship of physical parameters of ARs with flare activity in the context of the daily occurrence of C-, M-, and X-class flares. They used daily observations of solar full-disk line-of-sight magnetograms at 00:00 UT, and their models show limitations for the near-limb flares. Similarly, [33] assessed two additional attribution methods, DeepLIFT [34] and Integrated Gradients [35], for interpreting CNNs trained to predict solar flares. Their study used tracked AR patches within a range of $\pm 70^\circ$, where the interpretations are confined to central locations.

Furthermore, we presented deep learning-based full-disk flare prediction models trained with limited data in [15], [16] as preliminary studies on their feasibility for a complementary approach to operational forecasting systems. These models collectively showed relative success highlighting a promising direction for the development of more comprehensive full-disk models. However, a common limitation across both models was the lack of explainability. More recently, we presented explainable full-disk flare prediction models [19], [36], utilizing attribution methods to comprehend the models' effectiveness for near-limb flare events. We provided explanations for a limited set of near-limb flares through a case-based qualitative analysis. In this paper, we extend our approach beyond the

case-based qualitative analysis by employing a questionnaire-based evaluation scheme to quantitatively measure post hoc explanations for X-class flares in our dataset. This addition enhances our ability to show the efficacy of our model's predictions for both near-limb and central locations while also verifying its feasibility for operational forecasting systems.

III. DATA

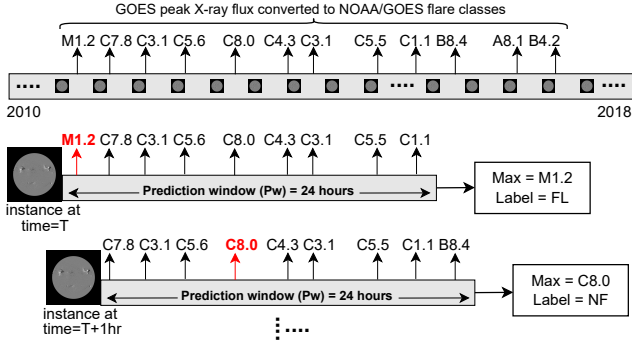


Fig. 1. A visualization for the data labeling procedure of hourly observations of full-disk line-of-sight magnetograms with a prediction window (Pw) of 24 hours. Here, ‘FL’ and ‘NF’ indicate the ‘Flare’ and ‘No Flare’ classes respectively in binary prediction mode ($\geq M$ -class flares).

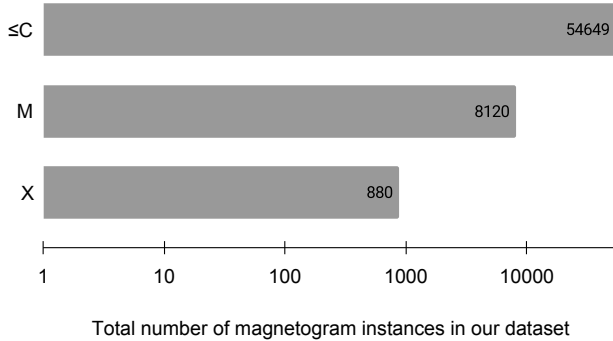


Fig. 2. The total number of hourly sampled magnetogram images per flare classes. Note that the length of the bar is in logarithmic scale.

We used full-disk line-of-sight solar magnetograms obtained from the Helioseismic and Magnetic Imager (HMI) [37] instrument onboard Solar Dynamics Observatory (SDO) [38] available as compressed JPEG 2000 (JP2) images in near real-time publicly via Helioviewer [39]. We sampled the magnetogram images every hour of the day, starting at 00:00 and ending at 23:00, from December 2010 to December 2018. We collected a total of 63,649 magnetogram images and labeled them using a 24-hour prediction window based on the maximum peak X-ray flux (converted to NOAA/GOES flare classes), as illustrated in Fig. 1. To elaborate, for a magnetogram instance at timestamp (T), if the maximum X-ray intensity of a flare was weaker than M within the next 24 hours, we labeled the observation as “No Flare” (NF: $<M$), and if it was $\geq M$, we labeled it as “Flare” (FL: $\geq M$). This resulted in 54,649 instances for the NF class

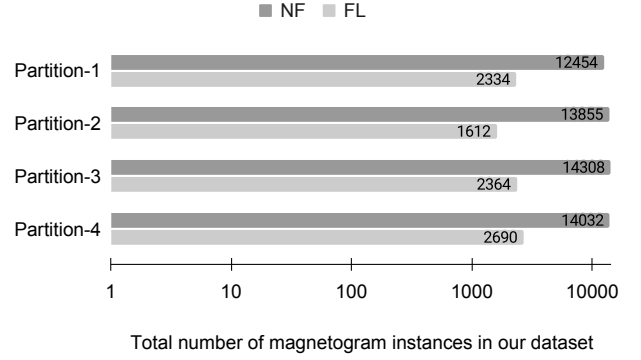


Fig. 3. Label distribution into four tri-monthly partitions for predicting $\geq M1.0$ -class flares. Note that the length of the bar is in logarithmic scale.

and 9,000 instances for the FL class. The detailed class-wise distribution of our data is shown in Fig. 2. Finally, we created a non-chronological split of our data into four temporally non-overlapping tri-monthly partitions introduced in [15] for our cross-validation experiments, as shown in Fig. 3. Since M- and X-class flares are relatively scarce events, for $\geq M$ -class flares the overall distribution of the data becomes highly imbalanced resulting in the FL to NF ratio in our dataset to be $\sim 1:6$.

IV. MODEL AND EXPLANATION

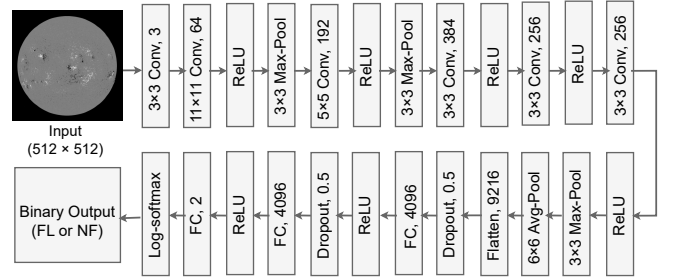


Fig. 4. The architecture of our AlexNet-based flare prediction model.

In this work, we utilized the AlexNet-based [40] full-disk flare prediction model in [19]. Our model extends the pre-trained AlexNet model to accommodate 1-channel input magnetogram images by introducing an additional convolutional layer at the beginning of the network that uses a 3×3 kernel, size-1 stride, and outputs three feature maps. This allows the network to take advantage of the pre-trained weights while processing the 1-channel magnetogram images. Moreover, the original AlexNet model is designed to handle 224×224 , 3-channel images, whereas our input images are 1-channel, 512×512 images. To address this discrepancy, an adaptive average pooling layer is used after the feature extraction process using the convolutional layer and before the fully-connected layer. This step ensures that the dimensions of the input images are matched to those expected by the pre-trained model. As a result, the model has six convolutional layers, three max pool layers, one adaptive average pool layer, and two fully connected layers. The architecture of the extended AlexNet model used in this study is illustrated in Fig. 4.

Deep learning models have become increasingly complex, creating difficulties in understanding the intricate data representations they learn. As a result, inconsistencies can arise in the patterns discovered by these models [41]. To address this, attribution methods have been proposed to aid in the interpretation of neural networks' decision-making process. Attribution methods use post hoc attention techniques to generate an attribution vector, or heat map, of the same size as the input, which visualizes how specific input sections contribute to the model's decision without influencing the decision-making process during model training and evaluation. There are mainly two categories of attribution methods: perturbation-based (e.g., Occlusion [42]) and gradient-based. Perturbation-based methods can lead to inconsistent interpretations due to the creation of Out-of-Distribution data caused by random perturbations [43], while gradient-based methods calculate the gradients of the output with respect to the extracted features or input using backpropagation, enabling attribution scores to be estimated more efficiently and robustly compared to input perturbations [44]. In this study, we employ a gradient-based method namely Guided Grad-CAM [31] to evaluate models and visualize their decisions, identifying the characteristics of magnetogram images that trigger specific decisions to help with operational forecasting under critical conditions.

The Guided Grad-CAM technique [31] is a fusion of the Grad-CAM and guided backpropagation [32] methods. Grad-CAM is a model-agnostic approach that does not require model retraining, and it employs the class-specific gradient data flowing into the final convolutional layer of a CNN to generate a rough localization map of the image's significant regions. Guided Backpropagation is based on the notion that neurons function as detectors for certain image characteristics. As a result, it computes the output gradient with respect to the input, but it only backpropagates the non-negative gradients through ReLU functions, emphasizing the pixels that are important in the image. Although the attributions generated by Grad-CAM are class-discriminative and localize relevant regions of the image, they do not highlight the pixel-level importance as accurately as guided backpropagation [45]. Guided Grad-CAM merges the globally precise details of guided backpropagation with the coarse localization advantages of Grad-CAM and is computed as the element-wise product of guided backpropagation with the upsampled Grad-CAM attributions. We used Captum [46] library to compute the Guided Grad-CAM attributions and the detailed steps used are enumerated as follows:

- 1) Load the pre-trained model and the desired input image X converted to a tensor. Run the forward pass of the model to obtain the final predictions y and the activations A from the last convolution layer.
- 2) Compute the gradient $\frac{dy_c}{dA_i}$ of the target class score y_c with respect to the activations A at location i . This shows how sensitive the class score is to changes in the activations.
- 3) Calculate the weight w_i for each activation location by averaging the gradients along the channels and spatial di-

mensions. Compute the Grad-CAM heatmap by performing a weighted sum of the activations followed by Rectified Linear Unit (ReLU) activation: $L_c = \text{ReLU}(\sum_i w_i \cdot A_i)$.

- 4) To compute the guided backpropagation mask:
 - a) Set the gradient of the target class score y_c with respect to the model output as 1, and all other gradients as 0.
 - b) Perform the backward pass through the network, calculating gradients for all activations and weights. During this process, all negative gradient values are set to 0 at each ReLU activation function. Hence, the backpropagation process propagates only positive gradients.
 - c) The output of the guided backpropagation will be a mask M , which highlights the pixels in the input that contribute positively to the target class score y_c .
- 5) Combine the Grad-CAM heatmap and guided backpropagation mask by multiplying L_c (upsampled to the input size) element-wise with M to obtain the Guided Grad-CAM heatmap $L'_c = L_c \cdot M$.

Finally, the Guided Grad-CAM heatmap L'_c is normalized and we can visualize it with or without overlaying on the original image, to identify the important regions contributing to the final prediction.

V. EXPERIMENTAL EVALUATION

In this section, we describe our experimental settings, including the augmentations used to balance the data, as well as the model implementation and hyperparameter configurations utilized in this study. Furthermore, we present the results and observations drawn from our experiments, with a particular focus on the locations of flares (near-limb flares), which are critical aspects of an operational forecasting system. In addition to reporting on the predictive performance of our models, we also present our findings on post-hoc explanations for X-class flares, which are the most relevant class of flares for Earth-impacting events. To quantify the quality of these explanations, we conducted a questionnaire-based assessment of their reliability, which is also described in this section.

A. Experimental Settings

The full-disk flare prediction model used in this study is trained using Stochastic Gradient Descent (SGD) as the optimizer and Negative Log-Likelihood (NLL) as the objective function. The model is initialized with pre-trained weights from the AlexNet Model [40], then further trained for 40 epochs with a batch size of 64 while employing a dynamic learning rate (initialized at 0.0099 and decreased by 5%). To overcome the issue of class imbalance, we utilized data augmentation and class weights in the loss function. Specifically, we used three augmentation techniques: vertical flipping, horizontal flipping, and rotations ranging from $+5^\circ$ to -5° . We applied augmentation to both classes, tripling the number of augmented samples for the entire FL class and randomly augmenting the NF class. After augmentation, we adjusted the class weights inversely proportional to the class frequencies to penalize misclassifications in the minority class.

The overall performance of our models is evaluated using two widely-used forecast skills scores: True Skill Statistics (TSS, shown in Eq. 1) and Heidke Skill Score (HSS, shown in Eq. 2), derived from the elements of confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). In the context of our solar flare prediction task, the FL class is considered as the positive outcome, while the NF class is negative.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \quad (1)$$

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

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

TSS and HSS values range from -1 to 1, where 1 indicates all correct predictions, -1 represents all incorrect predictions, and 0 represents no skill. In contrast to TSS, HSS is an imbalance-aware metric, and it is common practice to use HSS for the solar flare prediction models due to the high class-imbalance ratio present in the datasets. In solar flare prediction, TSS and HSS are the preferred evaluation metrics over commonly used ones in image classification, such as accuracy. This preference is because they offer a more thorough and dependable way to assess how well predictions work, especially when dealing with imbalanced class distributions. Lastly, we report the subclass and overall recall (shown in Eq. 3) for FL class instances (M-class and X-class) to demonstrate the prediction sensitivity in central and near-limb regions.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

To reproduce this work, the source code and experimental results can be accessed from our open-source repository [47].

B. Model Evaluation

The aggregated results from the 4-fold cross-validation experiments are as follows: Our models have on average $TSS \sim 0.51$ and $HSS \sim 0.38$, which improves over the performance of [15] by $\sim 4\%$ in terms of TSS (reported ~ 0.47) and $\sim 3\%$ in terms of HSS (reported ~ 0.35). In addition, on evaluating our results for correctly predicted (TP) and missed flare (FN) counts for class-specific flares (X-class and M-class) in central locations (within $\pm 70^\circ$) and near-limb locations (beyond $\pm 70^\circ$) of the Sun as shown in Table I, we observe that our models made correct predictions for $\sim 95\%$ of the X-class flares and $\sim 73\%$ of the M flares in central locations. Similarly, our models show a compelling performance for flares appearing on near-limb locations of the Sun, where $\sim 74\%$ of the X-class and $\sim 50\%$ of the M-class flares are predicted correctly¹. This is important because,

¹It is important to note that there are certain flares that may not be predicted at all with current observational capabilities, particularly when the active region responsible for the flare is not visible or only partially visible on the East limb up to 24 hours prior to the occurrence of the flare.

to our knowledge, the prediction of near-limb flares is often overlooked. More false positives in M-class are expected because of the model's inability to distinguish bordering class (C4+ to C9.9) flares from $\geq M1.0$ -class flares, which we have observed empirically in our prior work as well. Overall, we observed that $\sim 90\%$ and $\sim 66\%$ of the X-class and M-class flares, respectively, are predicted correctly by our models.

TABLE I
COUNTS OF CORRECTLY (TP) AND INCORRECTLY (FN) CLASSIFIED X- AND M-CLASS FLARES IN CENTRAL ($|longitude| \leq \pm 70^\circ$) AND NEAR-LIMB LOCATIONS. THE RECALL ACROSS DIFFERENT LOCATION GROUPS IS ALSO PRESENTED. COUNTS ARE AGGREGATED ACROSS FOLDS.

Flare-Class	Within $\pm 70^\circ$			Beyond $\pm 70^\circ$		
	TP	FN	Recall	TP	FN	Recall
X-Class	637	31	0.95	157	55	0.74
M-Class	4229	1601	0.73	1143	1147	0.50
Total (X&M)	4866	1632	0.75	1300	1202	0.52

We evaluated the effectiveness of our models both quantitatively and qualitatively by spatially analyzing their performance with respect to the locations of M- and X-class flares responsible for the labels. To conduct our analysis, we spatially binned the responsible flares based on their location in the Heliographic Stonyhurst (HGS) coordinate system, where each bin represents a 5° by 5° spatial cell in terms of latitude and longitude. Then, we analyzed the predictions of our models in the validation set from the 4-fold cross-validation experiments to determine whether the instances were correctly or incorrectly predicted. We calculated the recall separately for M-class, X-class, and both M- and X-class flares in each spatial cell to assess the models' sensitivity at a fine-grained level. The heatmaps that illustrate the spatial distribution of recall scores for our models can be found in Fig. 5. This allowed us to pinpoint the locations where our models were more effective in making accurate predictions and vice versa.

Our research findings indicate that our models are capable of effectively predicting even near-limb flares, particularly for X-class flares in both central and near-limb locations. However, we observed a higher number of false negatives in the near-limb areas for M-class flares. This represents a new capability in space weather forecasting, as we can now predict flares in regions where the magnetic field is distorted and pinpoint the locations of relevant active regions that are more likely to flare, which is critical for operational forecasting methods.

C. Evaluating Post hoc Explanations

In this section, we present a comprehensive analysis of the explanations generated using the Guided Grad-CAM attribution method for X-class flares. We focus on X-class flares because they are the most significant contributors to Earth-impacting extreme space weather events, such as the Halloween and Bastille Day Superstorms [48]. Our analysis encompasses both qualitative and quantitative evaluations of the generated explanations, providing a more detailed understanding of the attribution method's effectiveness in predicting

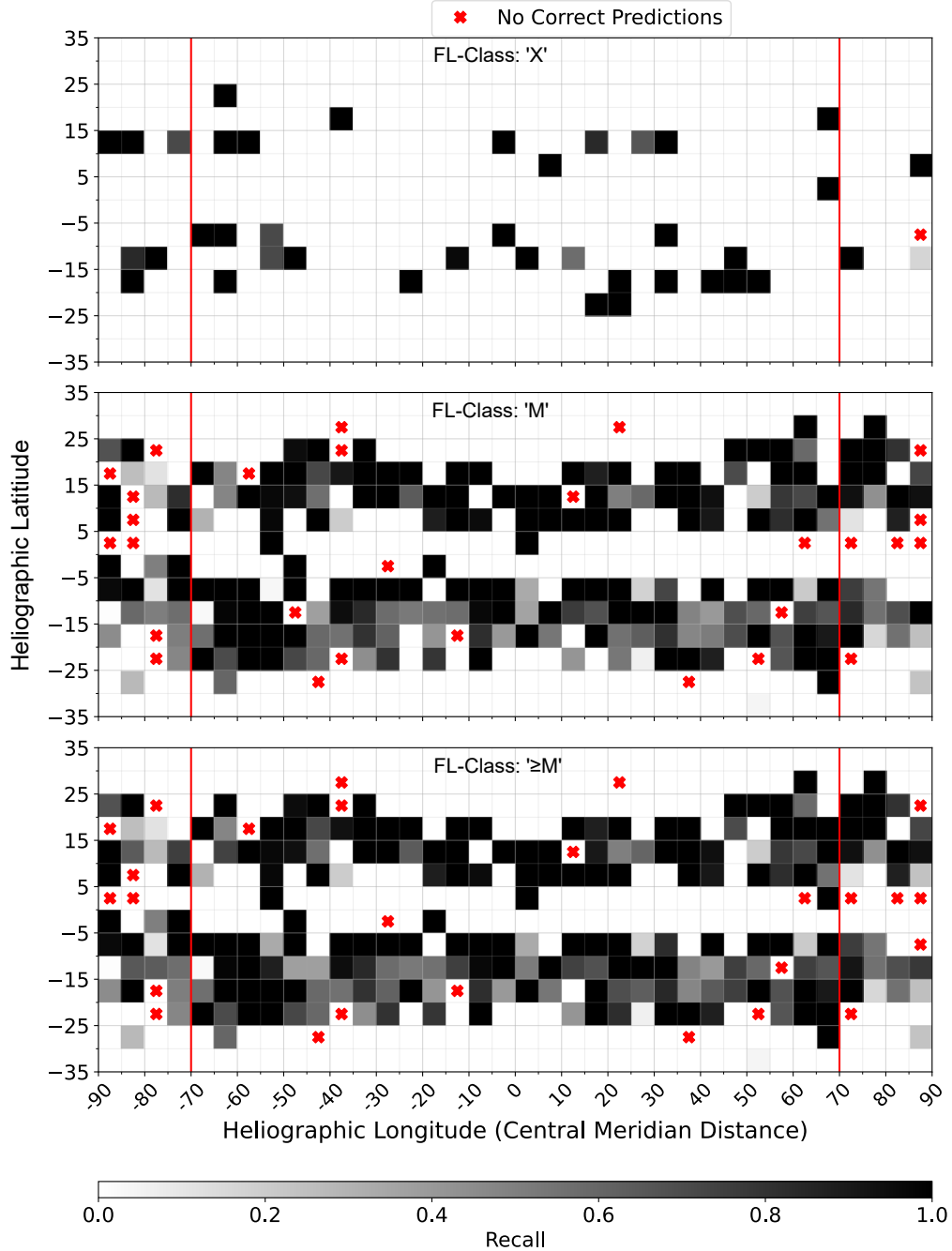


Fig. 5. A heatmap to assess the performance of our models in terms of recall for individual FL-Class (X- and M-class flares) and when combined ($\geq$ M-class flares) binned into $5^\circ \times 5^\circ$ flare locations used as the label. The flare events including and beyond $\pm 70^\circ$ longitude (separated by a vertical red line) are represented as near-limb events. Note: (i) Red cross in white grids represents locations with zero correct predictions while white cells without red cross represent unavailable instances. (ii) These results are aggregated from the validation sets of 4-folds.

X-class flares. By examining the results closely, we aim to improve our ability to forecast these powerful flares and mitigate their potential impacts.

For qualitative analytics, we evaluate the visual explanations with their temporal progression and interpret them by showcasing two examples of correctly predicted near-limb flares on both the East and West limbs of the Sun. Firstly, we present the explanation of our model for an East-limb X1.4-class (note that East and West are reversed in solar coordinates) flare

observed on 2011-09-22 at 10:29:00 UTC with visualizations generated using Guided Grad-CAM, denoted as GGCAM-MAP in Fig. 6 (a). For this, we used three input images, each separated by an hour starting at 2011-09-22 02:00:00 UTC (denoted as T), until the sunspot for the corresponding flare starts to become more visible in the magnetogram image. With temporal progression, we observed that as soon as a region becomes visible, the pixels covering the AR on the East limb are activated as shown in Fig. 6 (a), in fact, the model focuses

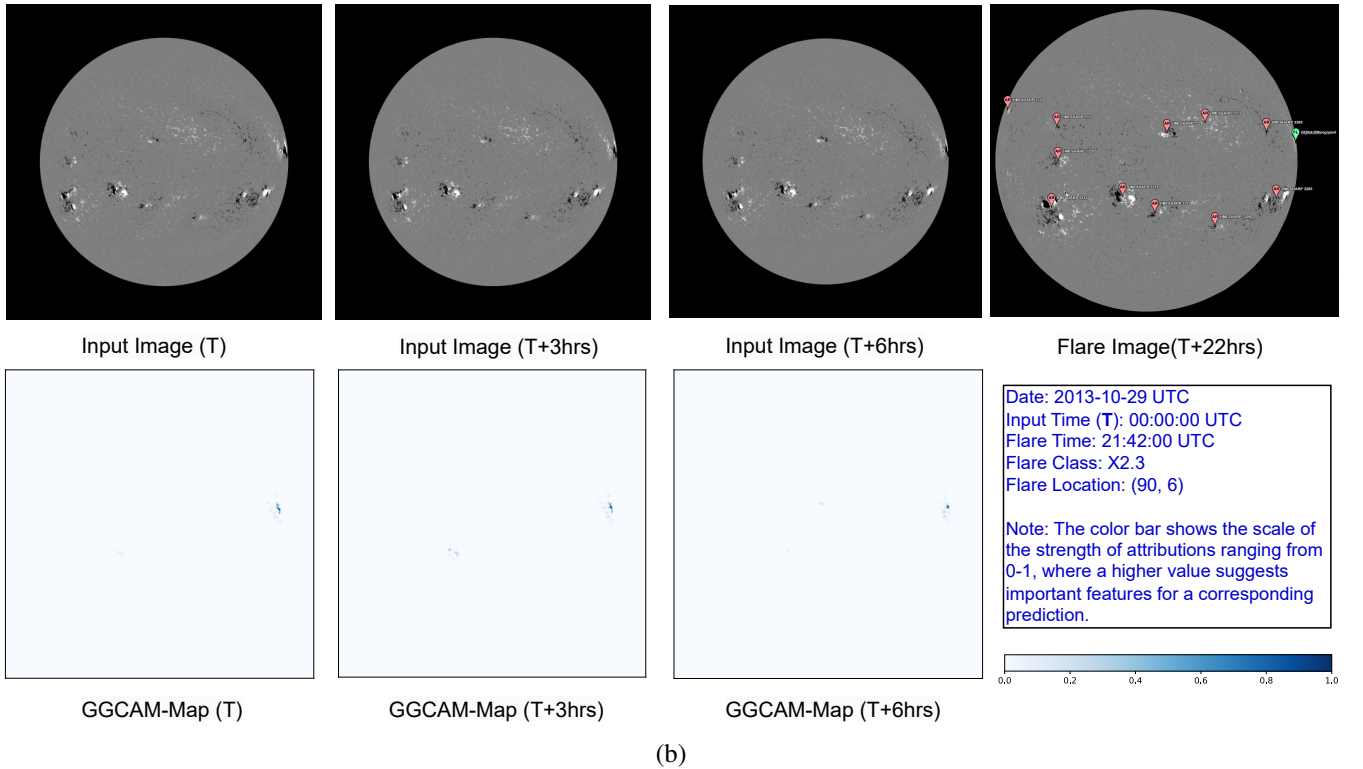
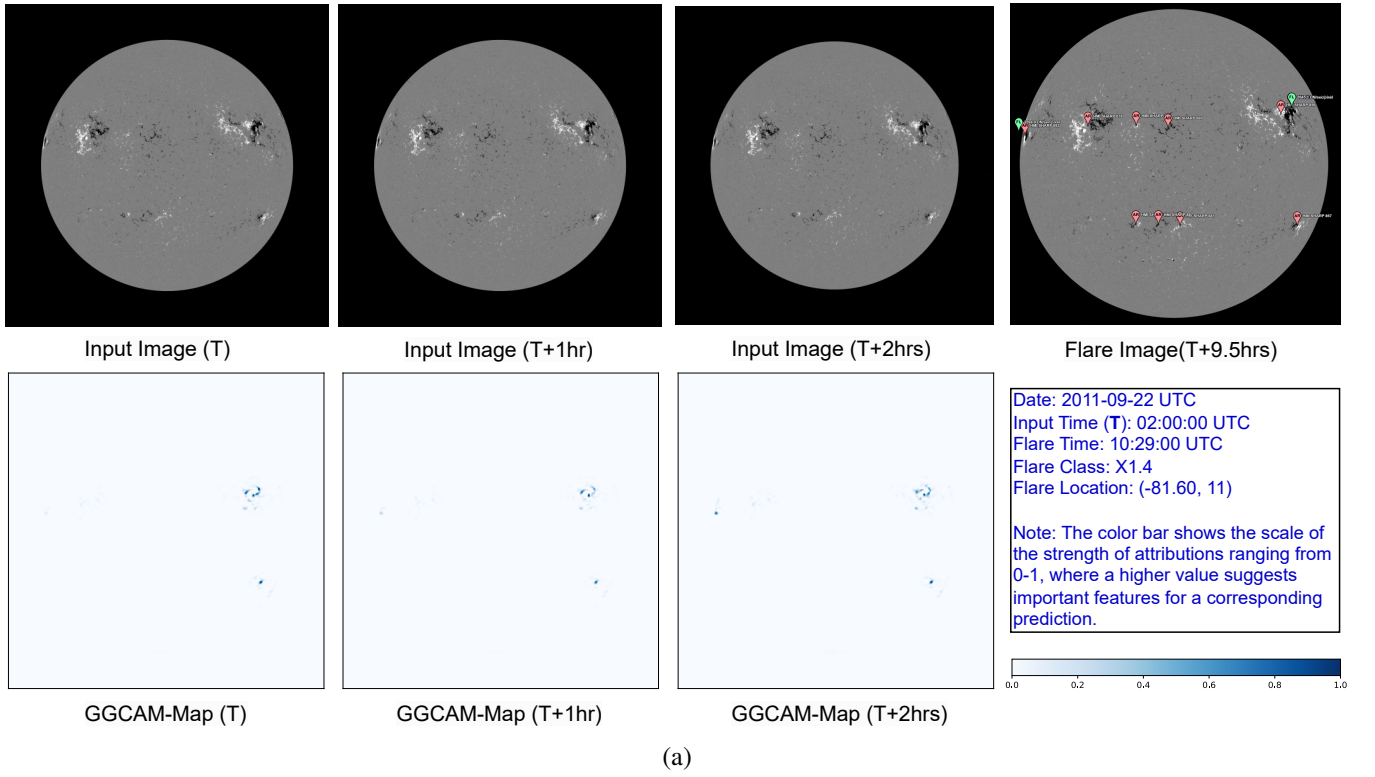


Fig. 6. A visual explanation of two correctly predicted near-limb X-class flares for East and West limbs, depicted in subfigures (a) and (b), respectively. The top row displays a series of three full-disk magnetogram input images, with a temporal progression of $i=\{0, 1, 2\}$ hours for (a) and $i=\{0, 3, 6\}$ hours for (b), starting from an initial timestamp T (different for each subfigure). These images are followed by a magnetogram image at approximately T+k hours, featuring overlays of detected active regions at the time of the flare ($k\sim 8.5$ hours in (a) and $k\sim 22$ hours in (b), denoted as Flare Image). Green flags indicate the flare location, while red flags represent all active regions in the magnetogram. The bottom row exhibits three activation maps generated using Guided Grad-CAM, corresponding to the input images and following the same temporal progression respective to subfigures (a) and (b).

TABLE II
TOTAL COUNTS OF GUIDED GRAD-CAM (TRUE WHEN ACTIVATED AND FALSE OTHERWISE) THAT CORRESPONDS TO THE FLARE LOCATION (FLOC: LOCATION OF MAXIMUM PEAK X-RAY FLUX USED TO LABEL EACH MAGNETOGRAM INSTANCES) FOR CORRECTLY (TP) AND INCORRECTLY (FN) CLASSIFIED X-CLASS FLARES IN CENTRAL ($|longitude| \leq \pm 70^\circ$) AND NEAR-LIMB LOCATIONS.

	Within $\pm 70^\circ$		Beyond $\pm 70^\circ$		Full-disk (Total)	
Prediction	True	False	True	False	True	False
TP	635	2	149	8	784	10
FN	31	0	55	0	86	0

on specific ARs including the relatively smaller area AR (in terms of pixels) on the East limb, even though other ARs are present in the magnetogram image. The visualization of attribution maps suggests that for this particular prediction, although only partially visible due to rotation and projection effects, the region responsible for the flare event is attributed to be important and hence contributes to the consequent decision.

Similarly, we analyze another case of correctly predicted near-limb flare (West-limb) of the Sun. For this, we provide a case of X2.3-class flare observed on 2013-10-29 at 21:42:00 UTC. For this, we again used three input images, each separated by three hours starting on 2013-10-29 at 00:00:00 UTC (denoted as T) shown in Fig. 6 (b). We observe that for the duration of six hours (until 16 hours prior to the flare event with three input images), the model’s activation for AR on the West-limb remains intact and it remains so as long as the AR corresponding to the flare is visible on the full-disk, indicating it is an important region for corresponding correct prediction. While there are several larger ARs present in the magnetogram, it focuses on specific ARs including the relatively smaller AR on the West limb. This shows that our models were able to learn shape-based features from compressed images and use them effectively for predicting flares, even in the presence of severe projection effects.

In addition, we conducted a rigorous quantitative evaluation of explanations for X-class flares, which comprised 880 instances in our dataset. We adopted a human-centered approach, as mentioned in [20], and designed a set of questionnaires addressing each instance to validate the explanations. The questionnaires were structured as follows:

- 1) Determine the activation of the location of the responsible X-class flare, denoted as FL, with maximum peak X-ray flux used to label the image. Assess the activation using a binary response format, with True (T) indicating activation and False (F) indicating non-activation. The location of FL is denoted as FLOC={T, F}.
- 2) Determine the activation of any other X-class flare locations besides the maximum flare location (FLOC) used to label the instance (i.e., FL) that coexisted within the 24-hour prediction window. Use a ternary response format, with ‘T’ indicating activation, ‘F’ indicating non-activation, and Not Applicable (N/A) indicating the absence of other X-class flares in the prediction window. This is denoted as XLOC={T, F, N/A}.

- 3) Determine the activation of any M-class flare locations that coexisted with FL within the 24-hour prediction window. Use a ternary response format, with ‘T’ indicating activation, ‘F’ indicating non-activation, and ‘N/A’ indicating the absence of M-class flares in the prediction window. This is denoted as MLOC={T, F, N/A}.
- 4) Determine the activation of any C-class flare locations that coexisted with FL within the 24-hour prediction window. Use a ternary response format, with ‘T’ indicating activation, ‘F’ indicating non-activation, and ‘N/A’ indicating the absence of C-class flares in the prediction window. This is denoted as CLOC={T, F, N/A}.

After examining the outcomes of question (1), which evaluated the activation of location for flare with maximum peak X-ray flux (FLOC) for X-class flares in our dataset, we found that FLOC was activated in 870 out of 880 instances, as presented in Table II. We subsequently conducted a separate evaluation of the outcomes for instances predicted correctly (TP) and incorrectly (FN) in both central and near-limb locations². Upon analyzing the results, we discovered that in every false negative (FN) instance, the model was able to locate the location of the flare but could not classify it accurately. Additionally, we observed that there were 10 instances ($\sim 1.1\%$ of total X-class flares), of which 8 were located in near-limb regions, while the remaining 2 were located in central locations, that were predicted correctly. However, the FLOC was not activated because other active regions were also activated, and those regions produced one or more M- and C-class flares. The location of the responsible flares was partially or completely invisible while in the near-limb region. Nevertheless, the model was able to predict these instances correctly.

In addition to question (1), we also incorporated the results of questions (2), (3), and (4) to examine the impact of other flares on the corresponding predictions, as presented in Table III. Notably, in instances where the FLOC is activated, we observed that only 23 instances had all other flare locations (i.e., XLOC, MLOC, and CLOC, if not N/A) activated. Furthermore, we found that all of the CLOC (if not N/A) were activated, even when the MLOC and FLOC were not activated. Overall, the majority of instances ($\sim 53\%$) showed activations for both MLOC and CLOC (if they coexist) along with FLOC. Our findings here point to a well-known phenomenon that flares productive active regions emit flares and eruptions in a clustered fashion [49], [50]. Moreover, we observed that, for almost 39% of the instances, where M-class locations were absent, only CLOC along with FLOC were activated. This may be due to the higher rates of occurrence of C-class flares, which constitute approximately 40% (25,150 instances) of the entire dataset and the model struggles to distinguish them from the FL-class as observed in our prior work [16]. Despite this, we observed the model’s adept ability to accurately identify

²Note that the central and near-limb locations referred to are only for the corresponding FL location (FLOC), and the evaluations are presented irrespective of the locations of other coexisting X, M, and C-class flares. Furthermore, the activations maps for incorrectly predicted instances (FN) are also generated and evaluated for positive class labels.

TABLE III

TOTAL COUNTS OF GUIDED GRAD-CAM THAT CORRESPONDS TO THE RESPONSIBLE FLARE LOCATION (FLOC), OTHER X-FLARES LOCATION (BESIDE MAXIMUM INTENSITY FLARE EVENT, I.E., FL-CLASS) DENOTED AS XLOC, OTHER M-FLARES LOCATION DENOTED AS MLOC, AND OTHER C-FLARES LOCATION DENOTED AS CLOC, WHICH COEXIST WITHIN THE PREDICTION WINDOW OF 24 HOURS. THE COUNT IS PRESENTED FOR CORRECTLY (TP) AND INCORRECTLY (FN) CLASSIFIED X-CLASS FLARES (FL) IN CENTRAL ($|longitude| \leq \pm 70^\circ$) AND NEAR-LIMB LOCATIONS.

When FL-location is activated in GGCAM (FLOC=T)			Within $\pm 70^\circ$		Beyond $\pm 70^\circ$		Full-disk (Total)	
Coexisting X-Flare Locations (XLOC)	Coexisting M-flare Locations (MLOC)	Coexisting C-Flare Locations (CLOC)	TP	FN	TP	FN	TP	FN
T	T	T	8	0	4	0	12	0
T	N/A	T	11	0	0	0	11	0
N/A	T	N/A	25	4	0	0	25	4
N/A	T	T	398	6	47	17	445	23
N/A	N/A	T	186	21	98	38	284	59
N/A	F	T	7	0	0	0	7	0
When FL-location is not activated in GGCAM (FLOC=F)								
N/A	T	T	0	0	6	0	6	0
N/A	N/A	T	0	0	1	0	1	0
N/A	F	T	2	0	1	0	3	0

relevant locations for the corresponding prediction, even for near-limb events, which emphasizes the utility of the full-disk model as a valuable contribution to operationalization efforts. Furthermore, it's important to note that the explanations show our model's ability to selectively highlight specific ARs that are more likely to initiate flares. This aspect is crucial for localizing flare-productive ARs, which is of utmost importance when performing long-term SEP event prediction [51].

VI. CONCLUSION AND FUTURE WORK

In this work, we used a gradient-based attribution method, namely, Guided Grad-CAM to interpret the predictions of our binary ($\geq M1.0$ -class) full-disk flare prediction model. We addressed the highly overlooked problem of flares appearing in near-limb locations of the Sun, and our evaluation shows a compelling performance of a full-disk model for such events. Furthermore, we evaluated our model's predictions with visual explanations and showed that the model's decisions are primarily based on characteristics corresponding to the relevant active regions in the magnetogram instances. As an extension, we plan to develop an automated method of evaluating explanations and incorporate the M-class flares as well to make this study more comprehensive. Furthermore, at this point, our models consider only the point-in-time spatial raster features in our data, and we intend to widen this work toward building models that can capture the temporal evolution of these features to improve predictive performance.

ACKNOWLEDGEMENTS

This project is supported in part under two NSF awards #2104004 and #1931555, jointly by the Office of Advanced Cyberinfrastructure within the Directorate for Computer and Information Science and Engineering, the Division of Astronomical Sciences within the Directorate for Mathematical and Physical Sciences, and the Solar Terrestrial Physics

Program and the Division of Integrative and Collaborative Education and Research within the Directorate for Geosciences. This work is also partially supported by the National Aeronautics and Space Administration (NASA) grant award #80NSSC22K0272. Data used in this study is a courtesy of NASA/SDO and the AIA, EVE, and HMI science teams and NOAA National Geophysical Data Center (NGDC).

REFERENCES

- [1] 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.
- [2] Y. Yasyukevich, E. Astafyeva, A. Padokhin, V. Ivanova, S. Sirovatkii, 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.
- [3] 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.
- [4] 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.
- [5] 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.
- [6] 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.
- [7] J. T. Hoeksema, Y. Liu, K. Hayashi, X. Sun, J. Schou, S. Couvidat, A. Norton, M. Bobra, R. Centeno, K. D. Leka, G. Barnes, and M. Turmon, "The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: Overview and performance," *Solar Physics*, vol. 289, no. 9, pp. 3483–3530, Mar. 2014.
- [8] M. B. Korsós, T. Baranyi, and A. Ludmány, "PRE-FLARE DYNAMICS OF SUNSPOT GROUPS," *The Astrophysical Journal*, vol. 789, no. 2, p. 107, Jun. 2014.
- [9] A. E. McCloskey, P. T. Gallagher, and D. S. Bloomfield, "Flaring rates and the evolution of sunspot group McIntosh classifications," *Solar Physics*, vol. 291, no. 6, pp. 1711–1738, Jun. 2016.

- [10] V. Deshmukh, T. E. Berger, E. Bradley, and J. D. Meiss, "Leveraging the mathematics of shape for solar magnetic eruption prediction," *Journal of Space Weather and Space Climate*, vol. 10, p. 13, 2020.
- [11] 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.
- [12] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, S. Watari, and M. Ishii, "Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms," *The Astrophysical Journal*, vol. 835, no. 2, p. 156, Jan. 2017.
- [13] 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.
- [14] 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.
- [15] 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)*. IEEE, Dec. 2021, pp. 1725–1730.
- [16] C. Pandey, R. Angryk, and B. Aydin, "Deep neural networks based solar flare prediction using compressed full-disk line-of-sight magnetograms," in *Information Management and Big Data*. Springer International Publishing, 2022, pp. 380–396.
- [17] K. Whitman, R. Egeland, I. G. Richardson, and et al., "Review of solar energetic particle models," *Advances in Space Research*, Aug. 2022.
- [18] J. Hong, A. Ji, C. Pandey, and B. Aydin, "Beyond traditional flare forecasting: A data-driven labeling approach for high-fidelity predictions," in *Big Data Analytics and Knowledge Discovery*. Springer Nature Switzerland, 2023, pp. 380–385.
- [19] 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," 2023. [Online]. Available: <https://arxiv.org/abs/2308.02682>
- [20] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, no. C, p. 89–106, Dec. 2021.
- [21] 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.
- [22] A. Devos, C. Verbeeck, and E. Robbrecht, "Verification of space weather forecasting at the regional warning center in Belgium," *Journal of Space Weather and Space Climate*, vol. 4, p. A29, 2014.
- [23] 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.
- [24] K. Leka, G. Barnes, and E. Wagner, "The NWRA classification infrastructure: description and extension to the discriminant analysis flare forecasting system (DAFFS)," *Journal of Space Weather and Space Climate*, vol. 8, p. A25, 2018.
- [25] 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.
- [26] M. B. Korsós, M. K. Georgoulis, N. Gyenge, S. K. Bisoi, S. Yu, S. Poedts, C. J. Nelson, J. Liu, Y. Yan, and R. Erdélyi, "Solar flare prediction using magnetic field diagnostics above the photosphere," *The Astrophysical Journal*, vol. 896, no. 2, p. 119, Jun. 2020.
- [27] M. G. Bobra and S. Couvidat, "Solar flare prediction using $\leq \text{SDO} \leq \text{HMI}$ VECTOR MAGNETIC FIELD DATA WITH a MACHINE-LEARNING ALGORITHM," *The Astrophysical Journal*, vol. 798, no. 2, p. 135, Jan. 2015.
- [28] 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, pp. 3549–3578, Apr. 2014.
- [29] S. Bhattacharjee, R. Alshehhi, D. B. Dhuri, and S. M. Hanasoge, "Supervised convolutional neural networks for classification of flaring and nonflaring active regions using line-of-sight magnetograms," *The Astrophysical Journal*, vol. 898, no. 2, p. 98, Jul. 2020.
- [30] K. Yi, Y.-J. Moon, D. Lim, E. Park, and H. Lee, "Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters," *The Astrophysical Journal*, vol. 910, no. 1, p. 8, Mar. 2021.
- [31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017.
- [32] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014.
- [33] Z. Sun, M. G. Bobra, X. Wang, Y. Wang, H. Sun, T. Gombosi, Y. Chen, and A. Hero, "Predicting solar flares using CNN and LSTM on two solar cycles of active region data," *The Astrophysical Journal*, vol. 931, no. 2, p. 163, Jun. 2022.
- [34] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2017.
- [35] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017.
- [36] C. Pandey, R. A. Angryk, and B. Aydin, "Explaining full-disk deep learning model for solar flare prediction using attribution methods," 2023. [Online]. Available: <https://arxiv.org/abs/2307.15878>
- [37] 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.
- [38] 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.
- [39] D. Muller, B. Fleck, G. Dimitoglou, B. Caplins, D. Amadiwe, J. Ortiz, B. Wamsler, A. Alexanderian, V. Hughitt, and J. Ireland, "JHelioviewer: Visualizing large sets of solar images using JPEG 2000," *Computing in Science & Engineering*, vol. 11, no. 5, pp. 38–47, Sep. 2009.
- [40] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014.
- [41] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, 2021.
- [42] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," 2013.
- [43] L. Qiu, Y. Yang, C. C. Cao, Y. Zheng, H. Ngai, J. Hsiao, and L. Chen, "Generating perturbation-based explanations with robustness to out-of-distribution data," in *Proceedings of the ACM Web Conference 2022*. ACM, Apr. 2022.
- [44] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 73–84, Jul. 2022.
- [45] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Mar. 2018.
- [46] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.
- [47] DMLab@GSU, "Open Source Repo: fdExplainGGCAM." [Online]. Available: <https://bitbucket.org/gsudmlab/fdexplainggcam/src/main/>
- [48] B. Tsurutani, A. Mannucci, B. Iijima, F. Guarnieri, W. Gonzalez, D. Judge, P. Gangopadhyay, and J. Pap, "The extreme halloween 2003 solar flares (and bastille day, 2000 flare), ICMes, and resultant extreme ionospheric effects: A review," *Advances in Space Research*, vol. 37, no. 8, pp. 1583–1588, Jan. 2006.
- [49] S. Toriumi and H. Wang, "Flare-productive active regions," *Living Reviews in Solar Physics*, vol. 16, no. 1, May 2019.
- [50] C. Campi, F. Benvenuto, A. M. Massone, D. S. Bloomfield, M. K. Georgoulis, and M. Piana, "Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence," *The ApJ*, vol. 883, no. 2, p. 150, Sep. 2019.
- [51] A. Ji, A. Arya, D. Kempton, R. Angryk, M. K. Georgoulis, and B. Aydin, "A modular approach to building solar energetic particle event forecasting systems," in *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2021, pp. 106–115.