Explainable Deep Learning-based Solar Flare Prediction with post hoc Attention for Operational Forecasting

Chetraj Pandey✉[1][0000−0002−4699−4050], Rafal A. Angryk[1][0000−0001−9598−8207], Manolis K. Georgoulis[2][0000−0001−6913−1330], and Berkay Aydin[1][0000−0002−9799−9265]

[1] Georgia State University, Atlanta, GA, USA
{cpandey1, rangryk, baydin2}@gsu.edu
[2] Research Center for Astronomy and Applied Mathematics, Academy of Athens, Athens, Greece
manolis.georgoulis@academyofathens.gr

Abstract. This paper presents a post hoc analysis of a deep learning-based full-disk solar flare prediction model. We used hourly full-disk line-of-sight magnetogram images and selected binary prediction mode to predict the occurrence of ≥M1.0-class flares within 24 hours. We leveraged custom data augmentation and sample weighting to counter the inherent class-imbalance problem and used true skill statistic and Heidke skill score as evaluation metrics. Recent advancements in gradient-based attention methods allow us to interpret models by sending gradient signals to assign the burden of the decision on the input features. We interpret our model using three post hoc attention methods: (i) Guided Gradient-weighted Class Activation Mapping, (ii) Deep Shapley Additive Explanations, and (iii) Integrated Gradients. Our analysis shows that full-disk predictions of solar flares align with characteristics related to the active regions. The key findings of this study are: (1) We demonstrate that our full disk model can tangibly locate and predict near-limb solar flares, which is a critical feature for operational flare forecasting, (2) Our candidate model achieves an average TSS=0.51±0.05 and HSS=0.38±0.08, and (3) Our evaluation suggests that these models can learn conspicuous features corresponding to active regions from full-disk magnetograms.

Keywords: Solar flares · Deep learning · xAI · Interpretability

1 Introduction

Solar flares are transient solar events of central importance to space weather forecasting, manifested as the sudden large eruption of electromagnetic radiation on the outermost atmosphere of the Sun. They are classified according to their peak X-ray flux level into the following five categories by National Oceanic and Atmospheric Administration (NOAA): X ($\geq 10^{-4} Wm^{-2}$), M ($\geq 10^{-5} Wm^{-2}$), C ($\geq 10^{-6} Wm^{-2}$), B ($\geq 10^{-7} Wm^{-2}$), and A ($\geq 10^{-8} Wm^{-2}$), where, X>M>C>B>A [6]. These flare classes are on a logarithmic scale, meaning that each class represents a tenfold increase in X-ray flux compared to the previous class. Large flares (M- and X-class) are scarce events that are more likely to incur a terrestrial impact and, therefore, the classes of interest that gather the attention of researchers. These flares may potentially disrupt the electricity supply chain, airline industry, and satellite communications, and pose radiation hazards to astronauts in space. To mitigate these risks, the necessity of a precise and reliable flare prediction model becomes imperative.

Active regions (ARs) on the Sun are places characterized by the largest accumulations of dipolar magnetic flux in the solar atmosphere. Most operational flare forecasts target these regions of interest and issue predictions for individual ARs, which are the main initiators of space weather events. To issue a full-disk forecast with an AR-based model, the output flare probabilities for each active region are usually aggregated using a heuristic function as mentioned in [20]. The heuristic function used to aggregate the final forecast operates under the assumption of conditional independence among ARs and that all ARs contribute equally to the aggregate forecast. This uniform weighting scheme may not accurately reflect the true influence of each AR on full-disk flare prediction probability. It is important to highlight that the weights of these ARs are generally unknown; there are no established methods to accurately determine them, nor are there any prior assumptions that guide the assignment of these weights.

Furthermore, the magnetic field measurements, employed by the AR-based forecasting techniques, are susceptible to severe projection effects as ARs get closer to limbs (to the degree that after ±60° the magnetic field readings are distorted [5]); therefore, the aggregated full-disk flare probability is in fact, restrictive (i.e., from ARs in central locations) as the data in itself is limited to ARs located within ±45° [11] to ±70° [9] and in some cases, even ±30° [8] due to severe projection effects [7]. As AR-based models include data up to ±70°, in the context of this paper, this upper limit (±70°) is used as a boundary for central (within ±70°) and near-limb regions (beyond ±70°).

In contrast to AR-based models, which use individual AR data from central locations, full-disk models use complete magnetogram images corresponding to the entire disk. These images are typically compressed JP2 (JPEG 2000) 8-bit representations (i.e., pixel values ranging from 0 to 255) derived from original magnetogram rasters which contain magnetic field strength values ranging from ∼±4500G. The compressed magnetogram images are used for shape-based parameters, e.g., size, directionality, borders, and inversion lines. Although projection effects still prevail in these images, full-disk models can learn from the near-limb areas. Thus, incorporating a full-disk model is essential to supplement AR-based models, enabling the prediction of flares in the Sun's near-limb areas and enhancing operational flare forecasting systems.

With recent advancements in machine learning and deep learning methods, their application in predicting solar flares has demonstrated great experimental success and accelerated the efforts of many interdisciplinary researchers [8,11,15,16,17,18,19,32]. Although deep learning methods have significantly enhanced solutions to image classification and computer vision problems, these models learn highly complex data representations, rendering them as black-box models. Consequently, the decision-making process within these models remains obscured, presenting a critical challenge for operational forecasting communities that rely on transparency to make informed decisions. Recently several empirical methods have been developed to explain and interpret the decisions made by deep neural networks. These are post hoc analysis methods (attribution methods) [12], meaning they focus on the analysis of trained models and do not contribute to the models' parameters while training. In this work, we primarily focus on developing a CNN-based full-disk model for solar flare prediction of ≥M1.0-class flares and evaluate and explain our model's performance by using three of the attribution methods: (i) Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) [25], (ii) Deep Shapley Additive Explanations (Deep SHAP) [13], and (iii) Integrated Gradients (IG) [31]. More specifically, we show that our model's decisions are based on the characteristics corresponding to ARs, and our models can tackle the flares appearing on near-limb regions.

The rest of this paper is organized as follows. In Sec. 2, we present the related work on flare forecasting. In Sec. 3, we present our methodology with data preparation and model architecture. In Sec. 4 we provide the description of all three post hoc explanation methods utilized in this work. In Sec. 5, we present our experimental settings, and model evaluation, and discuss the interpretation of our models, and in Sec. 6, we present our conclusions and future work.

2 Related Work

There have been several attempts to predict solar flares using machine learning and deep learning models. A multi-layer perceptron-based model was applied for solar flare prediction of ≥C1.0- and ≥M1.0-class flares in [15] by utilizing 79 manually selected physical precursors extracted from multi-modal solar observations. A CNN-based flare forecasting model trained with solar AR patches extracted from line-of-sight (LoS) magnetograms within ±30° of the central meridian to predict ≥C1.0-, ≥M1.0-, and ≥X1.0-class flares was presented in [8]. Similarly, [11] also used a CNN-based model to issue binary class predictions for both ≥C1.0- and ≥M1.0-class flares within 24 hours using AR patches located within ±45° of the central meridian. It is important to note that both of these models [8], [11] are limited to a small portion of the observable disk in central locations (within ±30° to ±45°) and thus possess the limited operational capability.

More recently, we presented a deep learning-based binary full-disk flare prediction model to predict ≥M1.0-class flares in [17] and to predict ≥C4.0- and ≥M1.0-class flares in [18] using bi-daily observations (i.e., two magnetograms per day) of full-disk LoS magnetograms. It is important to note that in [18] all the instances that fall between the ≥C4.0- and ≥M1.0-class flares were excluded in both training and validation sets. These particular sets of instances lie on the border of two binary outcomes and can be considered the harder-to-predict instances. These models are still black-box and do not provide explanations on any global and local variable importance. These explanations are important to understand the capabilities of full-disk models in near-limb regions and improve their trustworthiness in operational settings. In solar flare prediction, [2] used an occlusion-based method to interpret a CNN-based solar flare prediction model trained with AR patches. Similarly, [33] presented a deep learning-based flare prediction model for predicting C-, M-, and X-class flares and provided visual explanations using Grad-CAM [25], and Guided Backpropagation [28]. They used daily observations of solar full-disk LoS magnetograms at 00:00 UT, and their models show limitations for the near-limb flares. Moreover, in [30], DeepLIFT [27] and IG [31] were evaluated for explaining CNN-based flare prediction model trained using tracked AR patches within ±70°.

This paper presents a CNN-based model to predict ≥M1.0-class flares, trained with full-disk LoS magnetograms images. The novel contributions of this paper are as follows: (i) We show an improved overall performance of a full-disk solar flare prediction model, (ii) We utilized contemporary attribution methods to explain and interpret the decisions of our deep learning model, and (iii) More importantly, we show that our models can predict solar flares appearing on difficult-to-predict near-limb regions of the Sun.

3 Data and Model

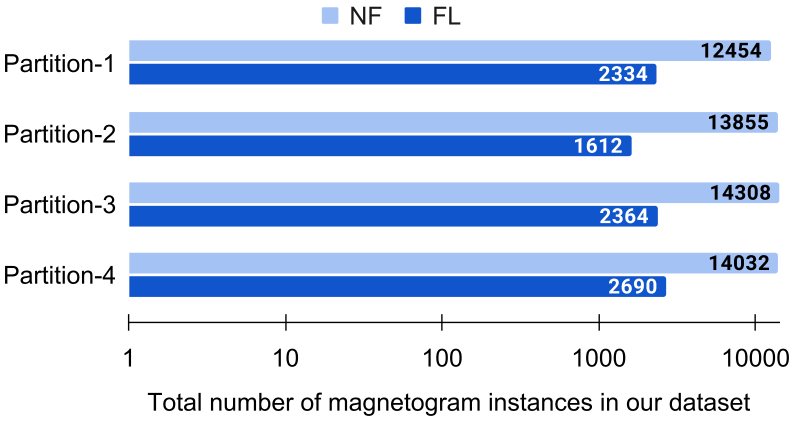


Fig. 1. Data distribution used in this study with four tri-monthly partitions for training ≥M1.0-class flare prediction models. Note: The length of the bars is in logarithmic scale.

We used full-disk LoS solar magnetograms obtained from the Helioseismic and Magnetic Imager (HMI) [24] instrument onboard Solar Dynamics Observatory (SDO) [21] available as compressed JP2 images in near real-time publicly via Helioviewer[1]. To enhance computational efficiency for training the deep learning model, these compressed images are resized to a smaller resolution of 512x512

[1] Helioviewer: `https://api.helioviewer.org`

pixels. We sampled hourly instances of magnetogram images at [00:00, 01:00, ..., 23:00] each day from Dec 2010 to Dec 2018. We labeled our data with a prediction window of 24 hours. The images are labeled based on the maximum peak X-ray flux (converted to NOAA flare classes) within the next 24 hours. We collect a total of 63,649 images and label them such that if the maximum X-ray intensity of flare is weaker than M1.0, the observations are labeled as "No Flare" (NF: <M1.0) and ≥M1.0 ones are labeled as "Flare" (FL: ≥M1.0). This results in 54,649 instances for the NF-class and 9,000 instances (8,120 instances of M-class and 880 instances of X-class flares) for the FL-class.

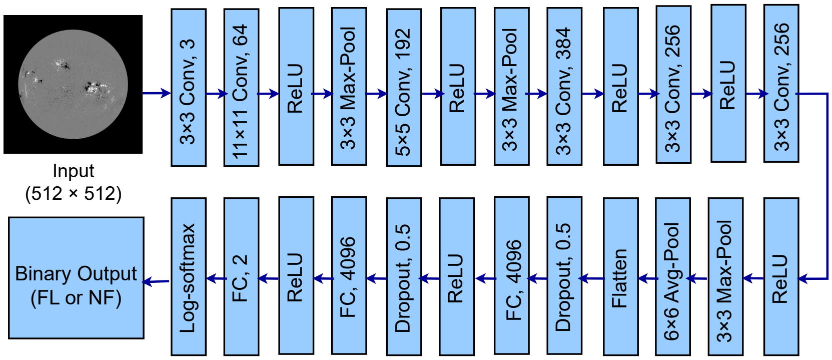


Fig. 2. The architecture of our full-disk flare prediction model.

We finally split our data into four temporally non-overlapping tri-monthly partitions for the cross-validation experiments. This partitioning of the dataset is created by dividing the data timeline from Dec 2010 to Dec 2018 into four partitions, where Partition-1 contains data from Jan to Mar, Partition-2 contains data from Apr to Jun, Partition-3 contains data from Jul to Sep, and finally, Partition-4 contains data from Oct to Dec as shown in Fig. 1. As a result of the infrequent occurrence of ≥M1.0-class flares, the dataset exhibits a significant imbalance, with the ratio of FL to NF class being approximately 1:6.

In this work, we extend the AlexNet [10] model by concatenating a convolutional layer at the beginning of the network to make use of the pre-trained weights for our 1-channel input magnetogram images as the pre-trained model requires a 3-channel image as input to the network. Our added convolutional layer uses a 3×3 kernel, size-1 stride, and outputs a 3-channel feature map which is then integrated into the standard AlexNet architecture as shown in Fig. 2. Furthermore, to efficiently utilize the pre-trained weights regardless of the architecture of the AlexNet model, which expects 224×224, 3-channel image as input, we use the adaptive average pooling after feature extraction before the fully-connected layer to match the dimension on our 1-channel, 512×512 magnetogram image. Overall, our model has six convolutional layers, three max-pool layers, one average-pool layer, and two fully-connected layers.

4 Interpretation Methods

Deep learning models are often deemed black-box due to their complex representations, resulting in interpretability, transparency, and consistency challenges concerning the patterns they learn [12]. To address this, various methods [34] have been proposed to interpret CNNs. One common approach is using attribution methods, which visualize how specific parts of the input influence the model's decisions. Attribution methods generate attribution vectors (heat maps) representing the contribution of each input element to the model's decision. These methods can be perturbation-based (e.g., Local Interpretable Model-Agnostic Explanations (LIME) [23]), involving altering the input and measuring the difference in output, or gradient-based, calculating gradients via backpropagation to estimate attribution scores. While perturbation-based methods suffer from inconsistency issues due to creating Out-of-Distribution data [22], gradient-based methods are more robust to input perturbations and computationally efficient [14]. Therefore, in this work, we employed three recent gradient-based methods to assess the interpretability of our models. By leveraging gradient-based techniques, known for their computational efficiency and robustness compared to perturbation-based methods, we aimed to visualize the decisions made by our model and gain insights into the specific characteristics in a magnetogram image that trigger the models' decisions. These methods allowed us to cross-validate and ensure the consistency of the explanations provided by our models, contributing to a more reliable and robust interpretation.

Guided Grad-CAM: The Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) method [25] leverages the benefits of the Grad-CAM and guided backpropagation [28]. Grad-CAM is a model-agnostic method that uses the class-specific gradient information flowing into the final convolutional layer of a CNN to produce a coarse localization map of the important regions in the image. Guided Backpropagation is based on the premise that the neurons act as detectors of certain image features, so it computes the gradient of the output with respect to the input, except that when propagating through ReLU functions, it only backpropagates the non-negative gradients and highlights the pixels that are important in the image. Attributions from Grad-CAM are class-discriminative and localize relevant image regions; however, do not highlight the fine-grained pixel importance as guided backpropagation [3]. Guided Grad-CAM combines the fine-grained details of guided backpropagation with the course localization advantages of Grad-CAM and is computed as the element-wise product of guided backpropagation with the upsampled Grad-CAM attributions.

Deep SHAP: SHAP values (SHapley Additive exPlanations) [13] is a method based on cooperative game theory[26] and used to increase the transparency and interpretability of machine learning models. SHAP shows the contribution of each feature to the prediction of the model, it does not evaluate the quality of the prediction itself. The contribution of each feature is calculated using cooperative game theory and Shapley values to assess how much each feature adds to the difference between the actual prediction and the average prediction. For deep-learning models, Deep SHAP [13] is considered an enhanced version of the DeepLIFT algorithm [27], where we approximate the conditional expectations of SHAP values using a selection of baseline samples from the dataset. The baselines typically contain a set of representative samples from the same distribution as the input data. For each input sample, it computes DeepLIFT attribution with respect to each baseline and averages resulting attributions. This method assumes that input features are independent of one another, and the explanations are modeled through the additive composition of feature effects.

Integrated Gradients: The last method we will analyze in this study is Integrated Gradients (IG) [31], which quantifies feature attributions by integrating the gradients of the model's output along a straight-line path from a baseline reference to the input feature under consideration. This method requires an extra input as the baseline, representing the non-appearance of the feature in the original image which is typically an all-zero vector. IG is favored for its completeness property, where the sum of integrated gradients for all features precisely equals the difference between the model's output for the given input and the baseline input values. This property ensures that the feature attributions accurately represent each feature's individual contribution to the model output, allowing us to reliably recover the model's output value by summing these contributions [29].

5 Experimental Evaluation

5.1 Experimental Settings

We trained a full-disk flare prediction model with stochastic gradient descent (SGD) as an optimizer and negative log-likelihood (NLL) as the objective function. Our model is initialized with pre-trained weights of AlexNet Model [10], and then we make use of a dynamic learning rate (initialized at 0.0099 and reduced 5%) to further train the model to 40 epochs with a batch size of 64. We address the class-imbalance issue using data augmentation and class weights to the loss function. We use three augmentation techniques: vertical flipping, horizontal flipping, and +5° to -5° rotations. We augment the data for both classes (where the entire FL-class data are augmented three times with three augmentation techniques and NF-class is augmented once randomly). We then adjust class weights inversely proportional to the class frequencies after augmentations. The use of class weights penalizes the misclassification made in the minority class. Our models are trained as 4-fold cross-validation experiments with each fold representing a different partition serving as the test set. Specifically, Fold-1 corresponds to Partition-1, Fold-2 corresponds to Partition-2, and so on.

We evaluate the performance of our models using two widely-used forecast skills scores: True Skill Statistics (TSS, in Eq. 1) and Heidke Skill Score (HSS, 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 paper, the FL class is the positive outcome and NF is the negative.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \tag{1}$$

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

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

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

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 in combination with TSS for the solar flare prediction models due to the high class-imbalance ratio present in the datasets. For a balanced dataset, these metrics are equivalent [1]. In solar flare prediction, TSS and HSS are the preferred choices of evaluation metrics compared to commonly used metrics in image classification (e.g., accuracy) as they ensure a comprehensive and reliable evaluation of predictive capabilities, especially in scenarios with imbalanced class distributions. Lastly, we report the subclass and overall recall (shown in Eq. 3) for flaring instances (M- and X-class) to assess the prediction sensitivity of our models in central and near-limb regions. To reproduce this work, the source code and experimental results can be accessed from our open-source repository [4].

5.2 Model Evaluation

Our models have on average TSS$\sim$0.51±0.05 and HSS$\sim$0.38±0.08, which improves over the performance of [17] by $\sim$4% in terms of TSS (reported 0.47±0.06) and by $\sim$3% in terms of HSS (reported 0.35±0.05) [2]. The detailed experimental results for each fold are shown in Table. 1.

In addition, we evaluate our results for correctly predicted and missed flare counts for class-specific flares (X-class and M-class) in central locations (within ±70°) and near-limb locations (beyond ±70°) of the Sun as shown in Table 2. We observe that our models made correct predictions for $\sim$95% of the X-class flares and $\sim$73% of the M-class 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 correctly. This is important because, 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 flares (C4.0 to C9.9) from ≥M1.0-class flares, which we have observed empirically in our prior work [18] as well.

[2] While there are several other works (mentioned in Sec. 2) in solar flare prediction, the results of these models are not directly comparable since they employ different datasets, data timelines, and data partitioning strategies.

Table 1. A comprehensive overview of 4-fold cross-validation experiments, showing all the four outcomes of confusion matrices (TP, FP, TN, FN) evaluated on the test sets, and performance of our models in terms of two skill scores (TSS and HSS).

Folds	TP	FP	TN	FN	TSS	HSS
Fold-1	1,720	1,943	10,511	614	0.58	0.47
Fold-2	1,155	3,083	10,772	457	0.49	0.29
Fold-3	1,585	2,668	11,640	779	0.48	0.36
Fold-4	1,706	2,241	11,791	984	0.47	0.40
Aggregated	6,166	9,935	44,714	2,834	**0.51±0.05**	**0.38±0.08**

Table 2. Counts of correctly (TP) and incorrectly (FN) classified X- and M-class flares in central ($|longitude| \leq \pm70^\circ$) and near-limb locations. The recall across different location groups is also presented. Counts are aggregated across folds.

	Within ±70°			Beyond ±70°		
Flare-Class	TP	FN	Recall	TP	FN	Recall
X-Class	637	31	0.95	157	55	0.74
M-Class	4,229	1,601	0.73	1,143	1,147	0.50
Total (X&M)	4,866	1,632	0.75	1,300	1,202	0.52

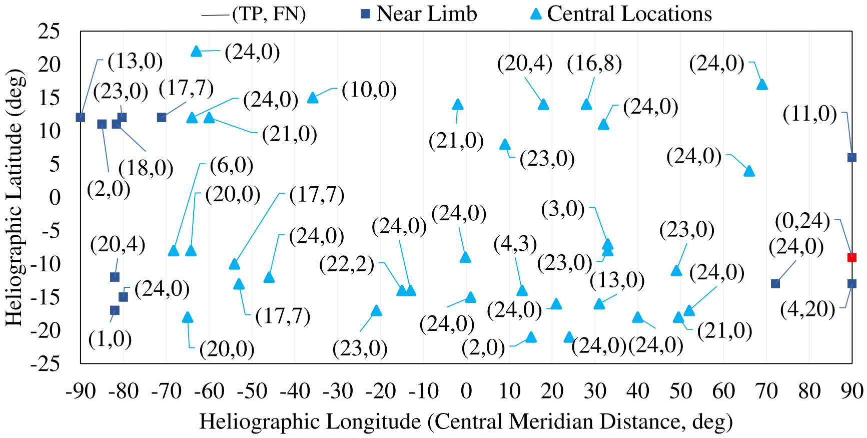


Fig. 3. A scatterplot to quantify the performance of our models in terms of True Positives (TP) and False Negatives (FN) for X-class flares grouped by flare locations. The flare events beyond ±70° longitude are represented as near-limb events. Note: (i) Red marker is for locations with zero TP. (ii) For some locations, TP+FN<24, given that we used hourly instances, is due to the unavailable instances from the source.

Overall, we observed that ~90% and ~66% of the X-class and M-class flares, respectively, are predicted correctly by our models.

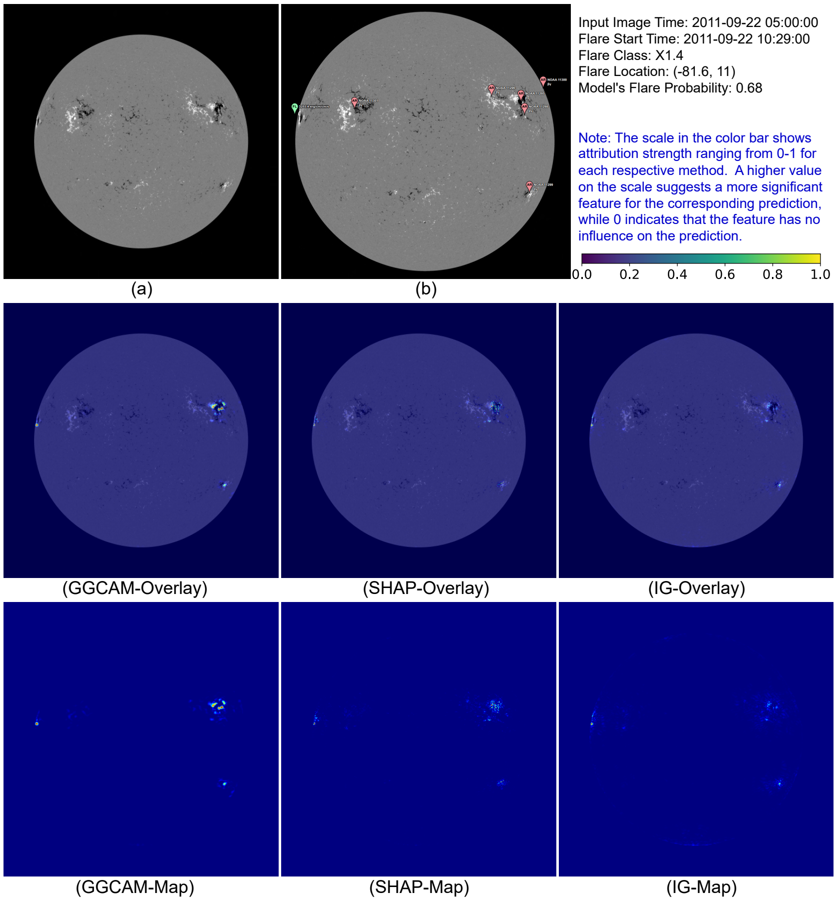
Furthermore, given that we sample our data with a 1-hour cadence resulting in 24 instances per day unless there are gaps due to unavailable data instances, any given flare instance is expected to be in the prediction window of 24 instances. X-class flares are relatively large flares that often dominate the prediction window. Therefore, we analyzed the predictions on X-class flares and observed that from a total of 45 X-class flare locations, our models correctly predict the occurrence of a flare at least once for 44 of them, as shown in Fig. 3. In particular, we show that the full-disk model presented in this paper can predict flares appearing on near-limb locations of the Sun at great accuracy, which provides a crucial addition to operational flare forecasting systems.

5.3 Model Interpretation

In this section, we present a case study, interpreting the visual explanations generated by our model, and also discuss the implications of these explanations in the operational forecasting scenario. For this, we use the visualizations generated using all three post hoc explanation methods mentioned earlier in Sec. 4 for two instances: (i) a correctly predicted (TP) near-limb flare instance and (ii) an incorrectly predicted (FP) instance.

Firstly, we interpret the predictions of our model for a correctly predicted X1.4-class flare observed on 2011-09-22 at 10:29:00 UTC on the East limb (note that East and West are reversed in solar coordinates). We generate a visual explanation using all three attribution methods. We utilized an input image from 2011-09-22 05:00:00 UTC (approximately 5.5 hours prior to the flare event) where the sunspot corresponding to the flare becomes visible in the magnetogram image. Interestingly, we observed that the pixels covering the AR on the East limb, which is responsible for the eventual X1.4 flare, are activated, as shown in Fig. 4. Note that the location of the flare is indicated by a green flag and all visible NOAA ARs are indicated by red flags in Fig. 4 (b). The model focuses on specific ARs, including the relatively smaller AR 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, the region responsible for the flare event is attributed as important, contributing to the consequent decision. This finding is consistent across all three methods, corroborating the explanation's reliability. However, Guided Grad-CAM and Deep SHAP provide finer details by suppressing noise compared to IG.

Similarly, to analyze a false positive case, we present an example of a C7.1 flare observed on 2014-01-06 at 00:08:00 UTC. To explain the result, we used an input magnetogram instance from 2014-01-05 06:00:00 UTC (∼18 hours prior to the event). The model's prediction probability for this instance being an FL-class is ∼0.97. Therefore, we seek a visual explanation of this prediction using all three interpretation methods. Upon analysis, we observed that the prediction mainly relies on only one AR, which indeed corresponds to the location of the eventual C7.1 flare (indicated by the green flag) when visualized with all three attribution methods, as shown in Fig. 5. This incorrect prediction can be attributed to the interference of the bordering class flares mentioned in [18]. Such


Input Image Time: 2011-09-22 05:00:00
Flare Start Time: 2011-09-22 10:29:00
Flare Class: X1.4
Flare Location: (-81.6, 11)
Model's Flare Probability: 0.68

Note: The scale in the color bar shows attribution strength ranging from 0-1 for each respective method. A higher value on the scale suggests a more significant feature for the corresponding prediction, while 0 indicates that the feature has no influence on the prediction.

(a) (b)

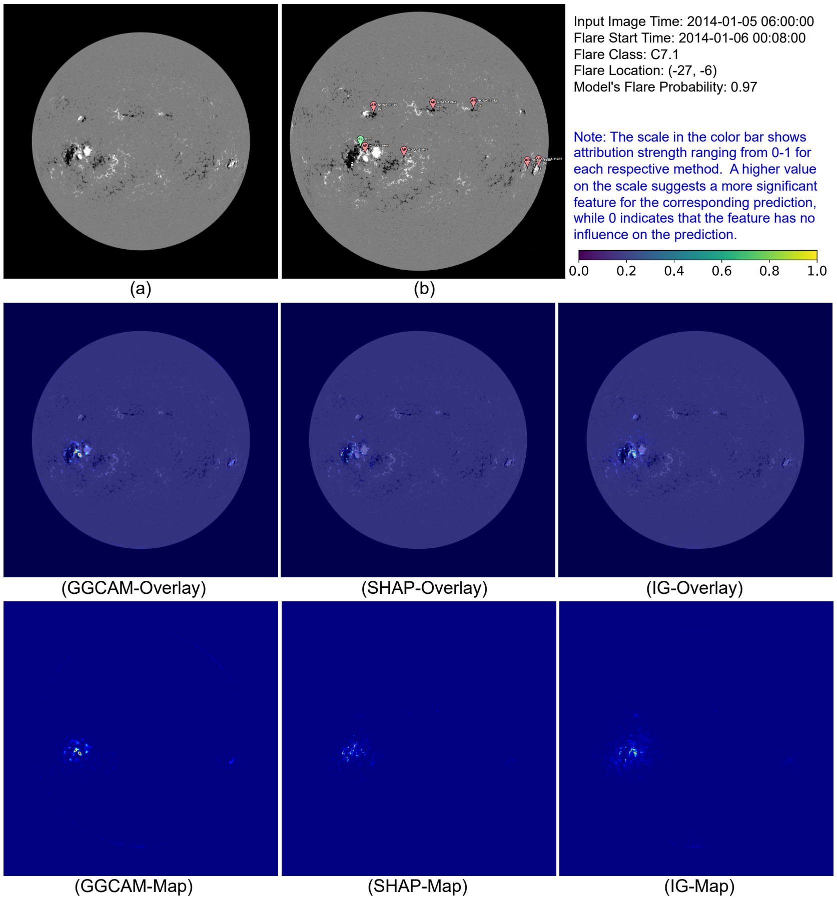
(GGCAM-Overlay) (SHAP-Overlay) (IG-Overlay)

(GGCAM-Map) (SHAP-Map) (IG-Map)

Fig. 4. A visual explanation for a correctly predicted near-limb FL-class instance. (a) Actual magnetogram from the dataset used as the input image. (b) Annotated full-disk magnetogram at flare start time, showing flare location (green flag) and NOAA ARs (red flags). Overlays (GGCAM, SHAP, IG) depict the input image overlayed with attributions, and Maps (GGCAM, SHAP, IG) showcase the attribution maps obtained from Guided Grad-CAM, Deep SHAP, and Integrated Gradients respectively.

interference poses a problem for binary flare prediction models. We noticed that out of 25,150 C-class flares, 9,240 flares led to incorrect predictions, accounting for approximately 37% of the total C-class flares in our dataset.

These two examples, although not exhaustive, carry significant implications for operational forecasting systems. By incorporating visual explanations into the forecasting process, in addition to providing a full-disk flare prediction probability, we have the capability to identify potential flare event locations among all visible ARs precisely. This is invaluable for improving the accuracy and reliability of solar flare forecasts, aiding in effective risk assessment and mitigation strategies. Furthermore, it provides a deeper understanding of the underlying fac-



Input Image Time: 2014-01-05 06:00:00
Flare Start Time: 2014-01-06 00:08:00
Flare Class: C7.1
Flare Location: (-27, -6)
Model's Flare Probability: 0.97

Note: The scale in the color bar shows attribution strength ranging from 0-1 for each respective method. A higher value on the scale suggests a more significant feature for the corresponding prediction, while 0 indicates that the feature has no influence on the prediction.

(a) (b)

(GGCAM-Overlay) (SHAP-Overlay) (IG-Overlay)

(GGCAM-Map) (SHAP-Map) (IG-Map)

Fig. 5. A visual explanation for an incorrectly predicted NF-class instance. (a) Actual magnetogram from the dataset used as the input image. (b) Annotated full-disk magnetogram at flare start time, showing flare location (green flag) and NOAA ARs (red flags). Overlays (GGCAM, SHAP, IG) depict the input image overlayed with attributions, and Maps (GGCAM, SHAP, IG) showcase the attribution maps obtained from Guided Grad-CAM, Deep SHAP, and Integrated Gradients respectively.

tors contributing to flare occurrences, empowering researchers and space weather experts to make more informed decisions and take timely actions to safeguard critical infrastructure and space assets.

6 Conclusion and Future Work

In this work, we used three recent gradient-based methods to interpret the predictions of our AlexNet-based binary flare prediction model trained for the prediction of ≥M1.0-class flares. We addressed the highly overlooked problem of flares appearing in near-limb locations of the Sun, and our model shows a compelling performance for such events. Furthermore, we evaluated our model's predictions with visual explanations, showing that the decisions are primarily capturing characteristics corresponding to the active regions in the magnetogram instance. Although our model shows improved capability, still suffers from high false positives attributed to high C-class flares. As an extension, we plan to study the individual class characteristics to obtain a better way of segregating these flare classes considering the background flux and generate a new set of labels that can better address the issue with border class flares. Furthermore, at this point, the models are only looking at the spatial patterns in our data, and we intend to widen this work toward spatiotemporal models to improve the 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. Ahmadzadeh, A., Aydin, B., Georgoulis, M., Kempton, D., Mahajan, S., Angryk, R.: How to train your flare prediction model: Revisiting robust sampling of rare events. The ApJ Supplement Series **254**(2), 23 (May 2021)
2. Bhattacharjee, S., Alshehhi, R., Dhuri, D.B., Hanasoge, S.M.: Supervised convolutional neural networks for classification of flaring and nonflaring active regions using line-of-sight magnetograms. The ApJ **898**(2), 98 (Jul 2020)
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: 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), `https://doi.org/10.1109/wacv.2018.00097`
4. DMLab: Source Code. `https://bitbucket.org/gsudmlab/explainingfulldisk/src/main/`
5. Falconer, D.A., Tiwari, S.K., Moore, R.L., Khazanov, I.: A new method to quantify and reduce the net projection error in whole-solar-active-region parameters measured from vector magnetograms. The ApJ **833**(2), L31 (Dec 2016)
6. Fletcher, L., Dennis, B.R., Hudson, H.S., Krucker, S., Phillips, K., Veronig, A., Battaglia, M., Bone, L., Caspi, A., Chen, Q., Gallagher, P., Grigis, P.T., Ji, H., Liu, W., Milligan, R.O., Temmer, M.: An observational overview of solar flares. Space Science Reviews **159**(1-4), 19–106 (Aug 2011)
7. Hoeksema, J.T., Liu, Y., Hayashi, K., Sun, X., Schou, J., Couvidat, S., Norton, A., Bobra, M., Centeno, R., Leka, K.D., Barnes, G., Turmon, M.: The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: Overview and performance. Solar Physics **289**(9), 3483–3530 (Mar 2014)

8. Huang, X., Wang, H., Xu, L., Liu, J., Li, R., Dai, X.: Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms. The ApJ **856**(1), 7 (Mar 2018), `https://doi.org/10.3847/1538-4357/aaae00`
9. Ji, A., Aydin, B., Georgoulis, M.K., Angryk, R.: All-clear flare prediction using interval-based time series classifiers. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 4218–4225. IEEE (Dec 2020)
10. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks (2014)
11. Li, X., Zheng, Y., Wang, X., Wang, L.: Predicting solar flares using a novel deep convolutional neural network. The ApJ **891**(1), 10 (Feb 2020)
12. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A review of machine learning interpretability methods. Entropy **23**(1), 18 (Dec 2020)
13. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 4768–4777. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
14. Nielsen, I.E., Dera, D., Rasool, G., Ramachandran, R.P., Bouaynaya, N.C.: Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. IEEE Signal Processing Magazine **39**(4), 73–84 (Jul 2022)
15. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Ishii, M.: Deep flare net (DeFN) model for solar flare prediction. The ApJ **858**(2), 113 (May 2018)
16. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M.: Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. The ApJ **835**(2), 156 (jan 2017)
17. Pandey, C., Angryk, R.A., Aydin, B.: Solar flare forecasting with deep neural networks using compressed full-disk HMI magnetograms. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 1725–1730. IEEE (Dec 2021), `https://doi.org/10.1109/bigdata52589.2021.9671322`
18. Pandey, C., Angryk, R.A., Aydin, B.: Deep neural networks based solar flare prediction using compressed full-disk line-of-sight magnetograms. In: Information Management and Big Data, pp. 380–396. Springer International Publishing (2022), `https://doi.org/10.1007/978-3-031-04447-2_26`
19. Pandey, C., Angryk, R.A., Aydin, B.: Explaining full-disk deep learning model for solar flare prediction using attribution methods (2023), `https://arxiv.org/abs/2307.15878`
20. Pandey, C., Ji, A., Angryk, R.A., Georgoulis, M.K., Aydin, B.: Towards coupling full-disk and active region-based flare prediction for operational space weather forecasting. Frontiers in Astronomy and Space Sciences **9** (Aug 2022), `https://doi.org/10.3389/fspas.2022.897301`
21. Pesnell, W., Thompson, B.J., Chamberlin, P.C.: The solar dynamics observatory (SDO). Solar Physics **275**(1-2), 3–15 (Oct 2011)
22. Qiu, L., Yang, Y., Cao, C.C., Zheng, Y., Ngai, H., Hsiao, J., Chen, L.: Generating perturbation-based explanations with robustness to out-of-distribution data. In: Proceedings of the ACM Web Conference 2022. ACM (Apr 2022), `https://doi.org/10.1145/3485447.3512254`
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (Aug 2016), `https://doi.org/10.1145/2939672.2939778`
24. Schou, J., Scherrer, P.H., Bush, R.I., Wachter, R., Couvidat, S., Rabello-Soares, M.C., Bogart, R.S., Hoeksema, J.T., Liu, Y., Duvall, T.L., Akin, D.J., Allard,

B.A., Miles, J.W., Rairden, R., Shine, R.A., Tarbell, T.D., Title, A.M., Wolfson, C.J., Elmore, D.F., Norton, A.A., Tomczyk, S.: Design and ground calibration of the helioseismic and magnetic imager (HMI) instrument on the solar dynamics observatory (SDO). Solar Physics **275**(1-2), 229–259 (Oct 2011)

25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV). IEEE (Oct 2017), https://doi.org/10.1109/iccv.2017.74
26. Shapley, L.: A Value for N-Person Games. RAND Corporation (1952)
27. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences (2019)
28. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net (2014), https://arxiv.org/abs/1412.6806
29. Sturmfels, P., Lundberg, S., Lee, S.I.: Visualizing the impact of feature attribution baselines. Distill **5**(1) (Jan 2020), https://doi.org/10.23915/distill.00022
30. Sun, Z., Bobra, M.G., Wang, X., Wang, Y., Sun, H., Gombosi, T., Chen, Y., Hero, A.: Predicting solar flares using CNN and LSTM on two solar cycles of active region data. The ApJ **931**(2), 163 (Jun 2022), https://doi.org/10.3847/1538-4357/ac64a6
31. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017), https://arxiv.org/abs/1703.01365
32. Whitman, K., Egeland, R., Richardson, I.G., Allison, C., Quinn, P., Barzilla, J., Kitiashvili, I., Sadykov, V., Bain, H.M., Dierckxsens, M., Mays, M.L., Tadesse, T., Lee, K.T., Semones, E., Luhmann, J.G., Núñez, M., White, S.M., Kahler, S.W., Ling, A.G., Smart, D.F., Shea, M.A., Tenishev, V., Boubrahimi, S.F., Aydin, B., Martens, P., Angryk, R., Marsh, M.S., Dalla, S., Crosby, N., Schwadron, N.A., Kozarev, K., Gorby, M., Young, M.A., Laurenza, M., Cliver, E.W., Alberti, T., Stumpo, M., Benella, S., Papaioannou, A., Anastasiadis, A., Sandberg, I., Georgoulis, M.K., Ji, A., Kempton, D., Pandey, C., Li, G., Hu, J., Zank, G.P., Lavasa, E., Giannopoulos, G., Falconer, D., Kadadi, Y., Fernandes, I., Dayeh, M.A., Muñoz-Jaramillo, A., Chatterjee, S., Moreland, K.D., Sokolov, I.V., Roussev, I.I., Taktakishvili, A., Effenberger, F., Gombosi, T., Huang, Z., Zhao, L., Wijsen, N., Aran, A., Poedts, S., Kouloumvakos, A., Paassilta, M., Vainio, R., Belov, A., Eroshenko, E.A., Abunina, M.A., Abunin, A.A., Balch, C.C., Malandraki, O., Karavolos, M., Heber, B., Labrenz, J., Kühl, P., Kosovichev, A.G., Oria, V., Nita, G.M., Illarionov, E., O'Keefe, P.M., Jiang, Y., Fereira, S.H., Ali, A., Paouris, E., Aminalragia-Giamini, S., Jiggens, P., Jin, M., Lee, C.O., Palmerio, E., Bruno, A., Kasapis, S., Wang, X., Chen, Y., Sanahuja, B., Lario, D., Jacobs, C., Strauss, D.T., Steyn, R., van den Berg, J., Swalwell, B., Waterfall, C., Nedal, M., Miteva, R., Dechev, M., Zucca, P., Engell, A., Maze, B., Farmer, H., Kerber, T., Barnett, B., Loomis, J., Grey, N., Thompson, B.J., Linker, J.A., Caplan, R.M., Downs, C., Török, T., Lionello, R., Titov, V., Zhang, M., Hosseinzadeh, P.: Review of solar energetic particle models. Advances in Space Research (Aug 2022), https://doi.org/10.1016/j.asr.2022.08.006
33. Yi, K., Moon, Y.J., Lim, D., Park, E., Lee, H.: Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters. The ApJ **910**(1), 8 (Mar 2021), https://doi.org/10.3847/1538-4357/abdebe
34. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (Mar 2021), https://doi.org/10.3390/electronics10050593