Explaining Full-disk Deep Learning Model for Solar Flare Prediction using Attribution Methods

Chetraj Pandey✉[0000−0002−4699−4050], Rafal A. Angryk[0000−0001−9598−8207], and Berkay Aydin[0000−0002−9799−9265]

Georgia State University, Atlanta, GA, 30303, USA
{cpandey1, rangryk, baydin2} @gsu.edu

Abstract. Solar flares are transient space weather events that pose a significant threat to space and ground-based technological systems, making their precise and reliable prediction crucial for mitigating potential impacts. This paper contributes to the growing body of research on deep learning methods for solar flare prediction, primarily focusing on highly overlooked near-limb flares and utilizing the attribution methods to provide a post hoc qualitative explanation of the model's predictions. We present a solar flare prediction model, which is trained using hourly full-disk line-of-sight magnetogram images and employs a binary prediction mode to forecast ≥M-class flares that may occur within the following 24-hour period. To address the class imbalance, we employ a fusion of data augmentation and class weighting techniques; and evaluate the overall performance of our model using the true skill statistic (TSS) and Heidke skill score (HSS). Moreover, we applied three attribution methods, namely Guided Gradient-weighted Class Activation Mapping, Integrated Gradients, and Deep Shapley Additive Explanations, to interpret and cross-validate our model's predictions with the explanations. Our analysis revealed that full-disk prediction of solar flares aligns with characteristics related to active regions (ARs). In particular, the key findings of this study are: (1) our deep learning models achieved an average TSS∼0.51 and HSS∼0.35, and the results further demonstrate a competent capability to predict near-limb solar flares and (2) the qualitative analysis of the model's explanation indicates that our model identifies and uses features associated with ARs in central and near-limb locations from full-disk magnetograms to make corresponding predictions. In other words, our models learn the shape and texture-based characteristics of flaring ARs even when they are at near-limb areas, which is a novel and critical capability that has significant implications for operational forecasting.

Keywords: Solar flares · Deep learning · Explainable AI.

1 Introduction

Solar flares are temporary occurrences on the Sun that can generate abrupt and massive eruptions of electromagnetic radiation in its outermost atmosphere. These events happen when magnetic energy, accumulated in the solar atmosphere, is suddenly discharged, leading to a surge of energy that spans a wide

range of wavelengths, from radio waves to X-rays. They are considered critical phenomena in space weather forecasting, and predicting solar flares is essential to understanding and preparing for their effects on Earth's infrastructure and technological systems. The National Oceanic and Atmospheric Administration (NOAA) classifies solar flares into five groups based on their peak X-ray flux level, namely A, B, C, M, and X, which represent the order of the flares from weakest to strongest [8] and are commonly referred to as NOAA/GOES flare classes, where GOES stands for Geostationary Operational Environmental Satellite. M- and X-class flares, which are rare but significant, are the strongest flares that can potentially cause near-Earth impacts, including disruptions in electricity supply chains, airline traffic, satellite communications, and radiation hazards to astronauts in space. This makes them of particular interest to researchers studying space weather. Therefore, developing better methods to predict solar flares is necessary to prepare for the effects of space weather on Earth.

Active regions (ARs) are typically characterized by strong magnetic fields that are concentrated in sunspots. These magnetic fields can become highly distorted and unstable, leading to the formation of plasma instabilities and the release of energy in the form of flares and other events [41]. Most operational flare forecasts target these regions of interest and issue predictions for individual ARs, which are the main initiators of space weather events. In order to produce a comprehensive forecast for the entire solar disk using an AR-based model, a heuristic function is used to combine the output flare probabilities ($P_{FL}(AR_i)$) for each active region (AR) [29]. The resulting probability, $P_{aggregated} = 1 - \prod_i \left[1 - P_{FL}(AR_i)\right]$, represents the likelihood of at least one AR producing a flare, assuming that the flaring events from different ARs are independent. However, there are two main issues with this approach for operational systems. Firstly, magnetic field measurements, which are the primary feature used by AR-based models, are subject to projection effects that distort measurements when ARs are closer to the limb. As a result, the aggregated full-disk flare probability is restricted to ARs in central locations, typically within $\pm 30^\circ$ [11], $\pm 45^\circ$ [20] to $\pm 70^\circ$ of the disk center [12]. Secondly, the heuristic function assumes that all ARs are equally important and independent of one another, which limits the accuracy of full-disk flare prediction probability. In contrast, full-disk models use complete magnetograms covering the entire solar disk, which are used to determine shape-based parameters such as size, directionality, borders, and inversion lines[13]. Although projection effects still exist in these images, full-disk models can learn from the near-limb areas and provide a complementary element to AR-based models by predicting flares that occur in these regions [28].

Machine learning and deep learning methods are currently being applied to predict solar flares, with experimental success and interdisciplinary collaboration from researchers in various fields [26], [25], [11], [20], [28], [14], [42]. Although these approaches have improved image classification and computer vision, they learn complex data representations, resulting in so-called black-box models. The decision-making process of these models is obscured, which is crucial for operational forecasting communities. To address this issue, several attribution meth-

ods, or post hoc analysis methods, have been developed to explain and interpret the decisions made by deep neural networks. These methods focus on analyzing trained models and do not contribute to the model's parameters during training. In this study, we develop a convolutional neural network (CNN) based full-disk model for predicting solar flares with a magnitude of ≥M-class flares. We evaluate and explain the model's performance using three attribution methods: Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) [32], Integrated Gradients [39], and Deep Shapley Additive Explanations (Deep SHAP) [22]. Our analysis reveals that our model's decisions are based on the characteristics corresponding to ARs, and it successfully predicts flares appearing on near-limb regions of the Sun.

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 a detailed description of all three attribution methods used as methods of explanation. In Sec. 5, we present our experimental evaluation. In Sec. 6, we discuss the interpretation of our models, and in Sec. 7, we present our conclusion and future work.

2 Related Work

Currently, there are four main types of methods in use for predicting solar flares, which include (i) human-based prediction techniques based on empirical observations [6], [7] (ii) statistical approaches [18], [19] (iii) numerical simulations based on physics-based models [17], [15], and (iv) data-driven models which made use of machine learning and deep learning techniques [4], [11], [20], [2], [28], [27]. The application of machine learning in predicting solar flares has seen significant progress due to recent advances. In one such application of machine learning, a multi-layer perceptron model based on machine learning was employed for predicting ≥C- and ≥M-class flares in [25] using 79 manually selected physical precursors derived from multi-modal solar observations.

Later, a CNN-based model was developed for predicting ≥C-, ≥M-, and ≥X-class flares using solar AR patches extracted from line-of-sight (LoS) magnetograms within ±30° of the central meridian in [11], taking advantage of the increasing popularity of deep learning models. [20] also used a CNN-based model to predict ≥C- and ≥M-class flares within 24 hours using AR patches located within ±45° of the central meridian. To address the class imbalance issue, they employed undersampling and data augmentation techniques. However, while undersampling led to higher experimental accuracy scores, it often failed to deliver similar real-time performance [1]. It is worth noting that both of these models have limited operational capability as they are restricted to a small portion of the observable disk in central locations (±30° and ±45°).

In addition, in [30], a CNN-based hybrid model was introduced which combined GoogleLeNet [40] and DenseNet [10]. The model was trained using a large volume of data from both the Helioseismic and Magnetic Imager (HMI) instrument onboard Solar Dynamics Observatory (SDO) and magnetograms from the

Michelson Doppler Imager (MDI) onboard the Solar and Heliospheric Observatory (SOHO). The aim of this model was to predict the occurrence of ≥C-class flares within the next 24 hours. However, it is important to note that these two instruments are not currently cross-calibrated for forecasting purposes, which may result in spurious or incomplete patterns being identified. More recently, an AlexNet-based [16] full-disk flare prediction model was presented in [28]. The authors provided a black-box model, but training and validation were limited due to a lower temporal resolution.

To interpret a CNN-based solar flare prediction model trained with AR patches, [3] used an occlusion-based method, and [43] presented visual explanation methods using daily observations of solar full-disk LoS magnetograms at 00:00 UT. They applied Grad-CAM [32] and Guided Backpropagation [36] to explore the relationship between physical parameters and the occurrence of C-, M-, and X-class flares. However, these methods had limitations in predicting near-limb flares. Recently, [38] evaluated two additional attribution methods, DeepLIFT [34] and Integrated Gradients [39], for interpreting CNNs trained on AR patches from central locations, i.e., within $\pm 70^{\circ}$ for predicting solar flares.

In this paper, a CNN-based model is presented for predicting ≥M-class flares, which was trained using full-disk LoS magnetogram images. The contributions of this study are threefold: (i) demonstrating an overall improvement in the performance of a full-disk solar flare prediction model, (ii) utilizing recent attribution methods to provide explanations of our model's decisions, and (iii) for the first time, demonstrating the capability of predicting flares in near-limb regions of the Sun, which are traditionally difficult to predict with AR-based models.

3 Data and Model

We used compressed images of full-disk LoS solar magnetograms obtained from the HMI/SDO available in near real-time publicly via Helioviewer[1] [23]. 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) within the next 24 hours, as illustrated in Fig. 1. To elaborate, if the maximum X-ray intensity of a flare was weaker than M (i.e., $< 10^{-5} Wm^{-2}$), we labeled the observation as "No Flare" (NF: <M), and if it was ≥M, we labeled it as "Flare" (FL: ≥M). This resulted in 54,649 instances for the NF class and 9,000 instances for the FL class. The detailed class-wise distribution of our data is shown in Fig. 2(a). Finally, we created a non-chronological split of our data into four temporally non-overlapping tri-monthly partitions for our cross-validation experiments. We created this partitioning by dividing the data timeline from December 2010 to December 2018 into four partitions. Partition-1 contained data from January to March, Partition-2 contained data from April to June,

[1] `HelioviewerAPIV2:https://api.helioviewer.org/docs/v2/`

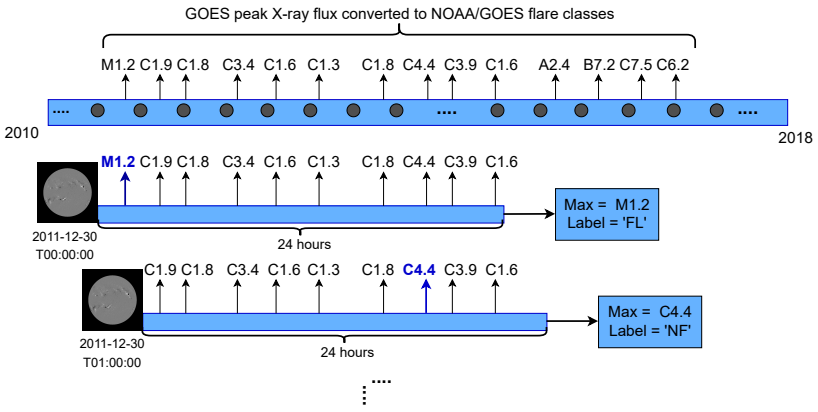

Fig. 1. A visual representation of the data labeling process using hourly observations of full-disk LoS magnetograms with a prediction window of 24 hours. Here, 'FL' and 'NF' indicates Flare and No Flare for binary prediction (≥M-class flares). The gray-filled circles indicate hourly spaced timestamps for magnetogram instances.

Partition-3 contained data from July to September, and Partition-4 contained data from October to December, as shown in Fig. 2(b). Due to the scarcity of ≥M-class flares, the overall distribution of the data is highly imbalanced, with FL:NF ∼1:6.

In our study, we employed transfer learning with a pre-trained VGG-16 model [35] for solar flare prediction. To use the pre-trained weights for our 1-channel input magnetogram images, we duplicated the channels twice, as the pre-trained model requires a 3-channel image for input. Additionally, we used the 7×7 adaptive average pooling after feature extraction using the convolutional layer and prior to the fully-connected layer to match the dimension of our 1-channel, 512×512 image. This ensures efficient utilization of the pre-trained weights, ir-

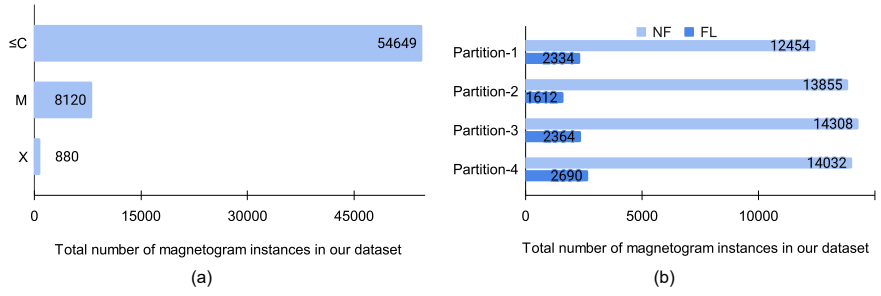

Fig. 2. (a) The total number of hourly sampled magnetograms images per flare classes. (b) Label distribution into four tri-monthly partitions for predicting ≥M-class flares.

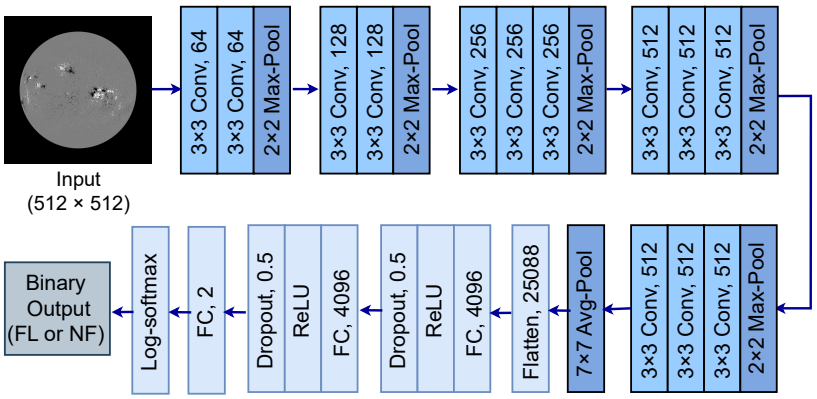

Fig. 3. The architecture of our full-disk solar flare prediction model.

respective of the architecture of the VGG-16 model, which is designed to receive 224×224, 3-channel images. Our model comprises 13 convolutional layers, each followed by a rectified linear unit (ReLU) activation, five max pool layers, one average pool layer, and three fully connected layers, as illustrated in Fig. 3.

4 Attribution Methods

Deep learning models are often seen as black boxes due to their intricate data representations, making them difficult to understand, leading to issues of inconsistency in the discovered patterns [21]. The attribution methods are post hoc approaches for model interpretation that provides insights into the decision-making process of the trained CNN models without influencing the training process. These methods generate an attribution vector, or heat map, of the same size as the input, where each element in the vector represents the contribution of the corresponding input element to the model's decision. Attribution methods can be broadly classified into two main categories: perturbation-based and gradient-based [9]. Perturbation-based methods modify the parts of the input to create new inputs and compute the attribution by measuring the difference between the output of the original and modified inputs. However, this approach can lead to inconsistent interpretations due to the creation of Out-of-Distribution (OoD) data caused by random perturbations [31]. In contrast, 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 to input perturbations [24].

Therefore, in this study, we employed three recent gradient-based methods to evaluate our models due to their reliability and computational efficiency. Our primary objective is to provide a visual analysis of the decisions made by our model and identify the characteristics of magnetogram images that trigger specific decisions by cross-validating the generated explanations from all three methods,

which can clarify the predictive output of the models and help with operational forecasting under critical conditions.

Guided Grad-CAM: The Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) method [32] combines the strengths of Grad-CAM and guided backpropagation [36]. Grad-CAM produces a coarse localization map of important regions in the image by using class-specific gradient information from the final convolutional layer of a CNN, while guided backpropagation calculates the gradient of the output with respect to the input, highlighting important pixels detected by neurons. While Grad-CAM attributions are class-discriminative and useful for localizing relevant image regions, they do not provide fine-grained pixel importance like guided backpropagation [5]. Guided Grad-CAM combines the fine-grained pixel details from guided backpropagation with the coarse localization advantages of Grad-CAM and generates its final localization map by performing an element-wise multiplication between the upsampled Grad-CAM attributions and the guided backpropagation output.

Integrated Gradients: Integrated Gradients (IG) [39] is an attribution method that explains a model's output by analyzing its features. To be more specific, IG calculates the path integral of gradients along a straight line connecting the baseline feature to the input feature in question. A baseline reference is required for this method, which represents the absence of a feature in the original image and can be a zero vector or noise; we used a zero vector of the size of the input as a baseline for our computation. IG is preferred for its completeness property, which states that the sum of integrated gradients for all features equals the difference between the model's output with the given input and the baseline input values. This property allows for attributions to be assigned to each individual feature and, when added together, should yield the output value itself [37].

Deep SHAP: SHAP values, short for SHapley Additive exPlanations [22], utilize cooperative game theory [33] to enhance the transparency and interpretability of machine learning models. This method quantifies the contribution or importance of each feature on the model's prediction rather than evaluating the quality of the prediction itself. In the case of deep-learning models, Deep SHAP [22] improves upon the DeepLIFT algorithm [34] by estimating the conditional expectations of SHAP values using a set of background samples. For each input sample, the DeepLIFT attribution is computed with respect to each baseline, and the resulting attributions are averaged. This method assumes feature independence and explains the model's output through the additive composition of feature effects. Although it assumes a linear model for each explanation, the overall model across multiple explanations can be complex and non-linear. Similar to IG, Deep SHAP also satisfies the completeness property [37].

5 Experimental Evaluation

5.1 Experimental Settings

We trained a full-disk flare prediction model using Stochastic Gradient Descent (SGD) as an optimizer and Negative Log-Likelihood (NLL) as the objective

function. To apply NLL loss, we used logarithmic-softmax activation on the raw logits from the output layer. Our model was initialized with pre-trained weights from the VGG-16 model [35]. We further trained the model for 50 epochs with a batch size of 64 using dynamic learning rates (initialized at 0.001 and halved every 5 epochs). To address the class imbalance issue, we used data augmentation and class weights in the loss function. Specifically, we applied three augmentation techniques (vertical flipping, horizontal flipping, and rotations of $+5^\circ$ to -5°) during the training phase to explicitly augment the minority FL-class three times. However, this still left the dataset imbalanced, so we adjusted the class weights inversely proportional to the class frequencies after augmentations and penalized misclassifications made in the minority class. To improve the generalization of our model without introducing bias in the test set, we applied data augmentation exclusively during the training phase, and we opted for augmentation over oversampling and undersampling as the latter two may lead to overfitting of the model [2]. Finally, we conducted 4-fold cross-validation experiments using tri-monthly partitions to train our models.

We assess the overall performance of our models using two 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 Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In this context, FL and NF represent positive and negative classes respectively.

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

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

where N = TN + FP and P = TP + FN. 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 and for a balanced dataset, these metrics are equivalent as discussed in [1]. Lastly, we report the subclass and overall recall for flaring instances (M- and X-class), which is calculated as $(\frac{TP}{TP+FN})$, to demonstrate the prediction sensitivity. To reproduce this work, the source code and detailed experimental results can be accessed from our open source repository [2].

5.2 Evaluation

We performed 4-fold cross-validation using the tri-monthly dataset for evaluating our models. Our models have on average TSS∼0.51 and HSS∼0.35, which improves over the performance of [28] by ∼4% in terms of TSS (reported ∼0.47)

[2] `explainFDvgg16:https://bitbucket.org/gsudmlab/explainfdvgg16/src/main/`

and competing results in terms of HSS (reported ∼0.35). 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 1. We observe that our models made correct predictions for ∼89% of the X-class flares and ∼77% 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 ∼77% of the X-class and ∼52% 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 [C4+ to C9.9] flares from ≥M-class flares, which we have observed empirically in our prior work [27] as well. Overall, we observed that ∼86% and ∼70% of the X-class and M-class flares, respectively, are predicted correctly by our models.

We also quantitatively and qualitatively evaluated our models' effectiveness 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 have spatially binned the responsible flares (maximum X-ray flux within the next 24h) and analyzed whether these instances were correctly (TP) or incorrectly predicted (FN). For this, we used the predictions of our models in the validation set from the 4-fold cross-validation experiments. Here, each bin represents a 5° by 5° spatial cell in Heliographic Stonyhurst (HGS) coordinate system (i.e., latitude and longitude). For each subgroup, represented in a spatial cell, we calculate the recall for M-class, X-class, and M- and X-class flares, separately to assess the models' sensitivity at a fine-grained level. The heatmaps demonstrating the spatial distribution of recall scores of our models can be seen in Fig. 4. This allows us to pinpoint the locations where our models were more effective in making accurate predictions and vice versa. We observed that our models demonstrated reasonable performance overall, particularly for X-class flares, in both near-limb and central locations. However, we also observed a higher number of false negatives around near-limb locations for M-class flares. In particular, we demonstrate that the full-disk model proposed in this paper can predict flares appearing at

Table 1. Counts of correctly (TP) and incorrectly (FN) classified X- and M-class flares in central ($|longitude| \leq \pm 70°$) 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	597	71	0.89	164	48	0.77
M-Class	4,464	1,366	0.77	1,197	1,093	0.52
Total (X&M)	5,061	1,437	0.78	1,361	1,141	0.54

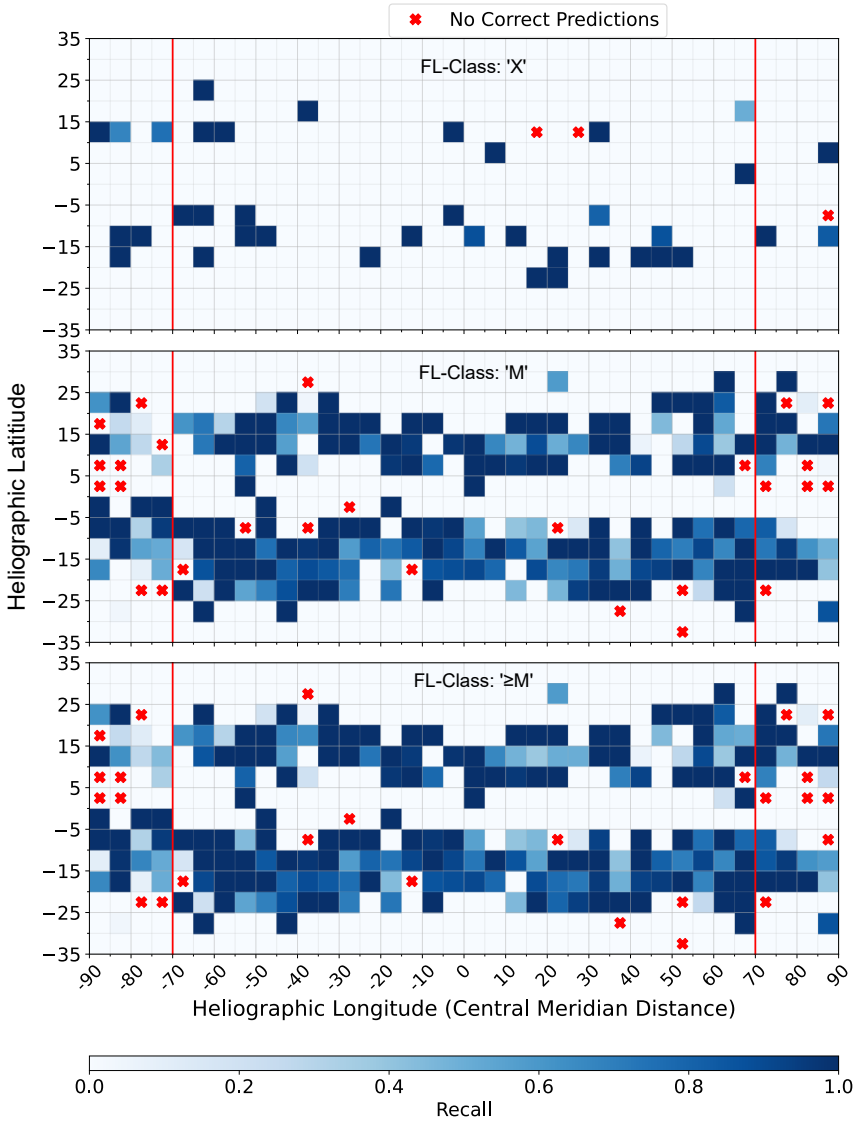

Fig. 4. A heat map showcasing recall for individual FL-Class (X- and M-class flares) and when combined (≥M-class flares) binned into 5° × 5° flare locations used as the label. The flare events beyond ±70° longitude (separated by a vertical red line) represent near-limb events. Note: Red cross in white grids represents locations with zero correct predictions while white cells without red cross represent unavailable instances.

near-limb locations of the Sun with great accuracy, which is a crucial addition to operational forecasting systems.

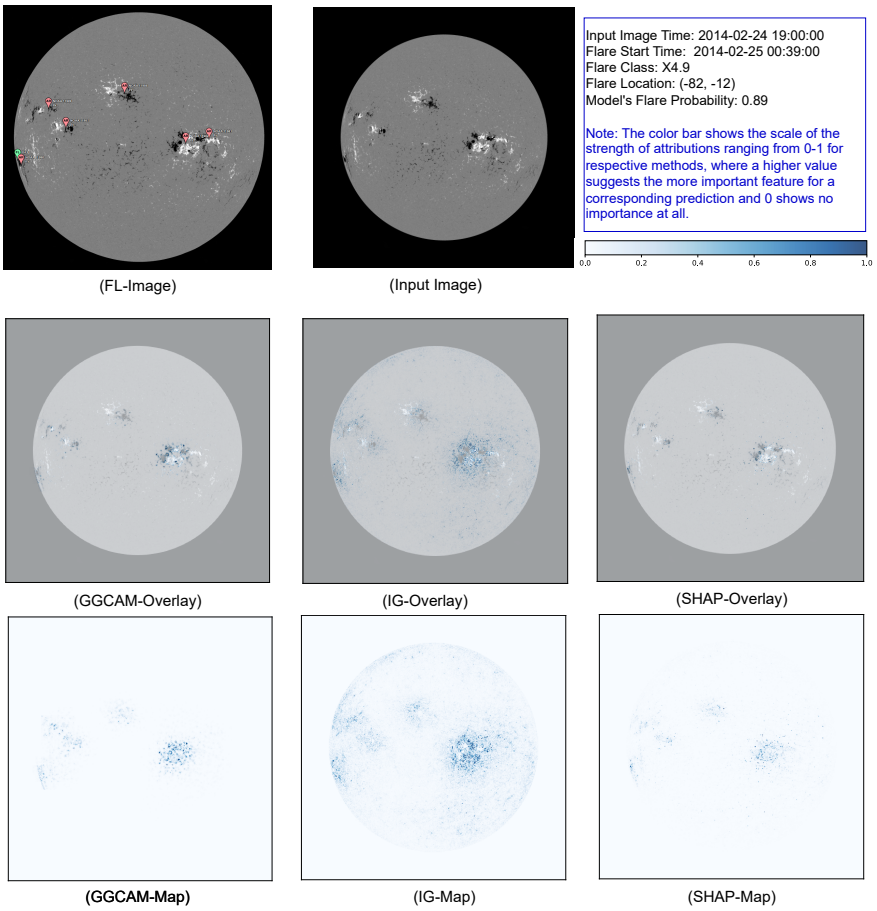

Fig. 5. A visual explanation of correctly predicted near-limb (East) FL-class instance. (FL-Image): Annotated full-disk magnetogram at flare start time, showing flare location (green flag) and NOAA ARs (red flags). (Input Image): Actual magnetogram from the dataset. Overlays (GGCAM, IG, SHAP) depict the input image overlayed with attributions, and Maps (GGCAM, IG, SHAP) showcase the attribution maps obtained from Guided Grad-CAM, Integrated Gradients, and Deep SHAP, respectively.

6 Discussion

In this section, we interpret the visual explanations generated using the attribution methods mentioned earlier for correctly predicted near-limb flares and the model's high confidence in an incorrect prediction. As the major focus of this study is on the near-limb flares, we interpret the predictions of our model for an east-limb X4.9-class (note that East and West are reversed in solar coordinates) flare observed on 2014-02-25 at 00:39:00 UTC with a visual explanation generated using all three attribution methods. For this, we used an input im-

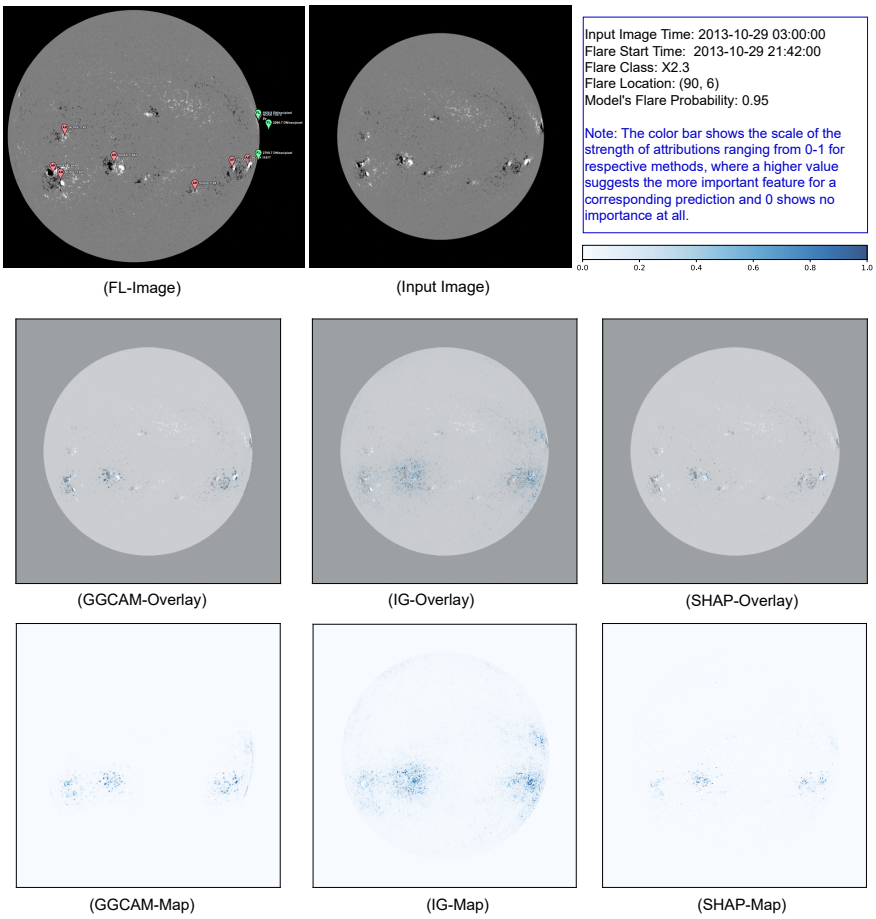

Fig. 6. A visual explanation of correctly predicted near-limb (West) FL-class instance. (FL-Image): Annotated full-disk magnetogram at flare start time, showing flare location (green flag) and NOAA ARs (red flags). (Input Image): Actual magnetogram from the dataset. Overlays (GGCAM, IG, SHAP) depict the input image overlayed with attributions, and Maps (GGCAM, IG, SHAP) showcase the attribution maps obtained from Guided Grad-CAM, Integrated Gradients, and Deep SHAP, respectively.

age at 2014-02-24 19:00:00 UTC (∼6 hours prior to the flare event), where the sunspot for the corresponding flare becomes visible in the magnetogram image. We observed while all three methods highlight features corresponding to an AR in the magnetogram, Guided Grad-CAM and Deep SHAP provide finer details by suppressing noise compared to IG as shown in Fig. 5. Furthermore, the visualization of attribution maps suggests that for this particular prediction, although barely visible, the region responsible for the flare event is considered important and hence contributes to the consequent decision. The explanation shows that

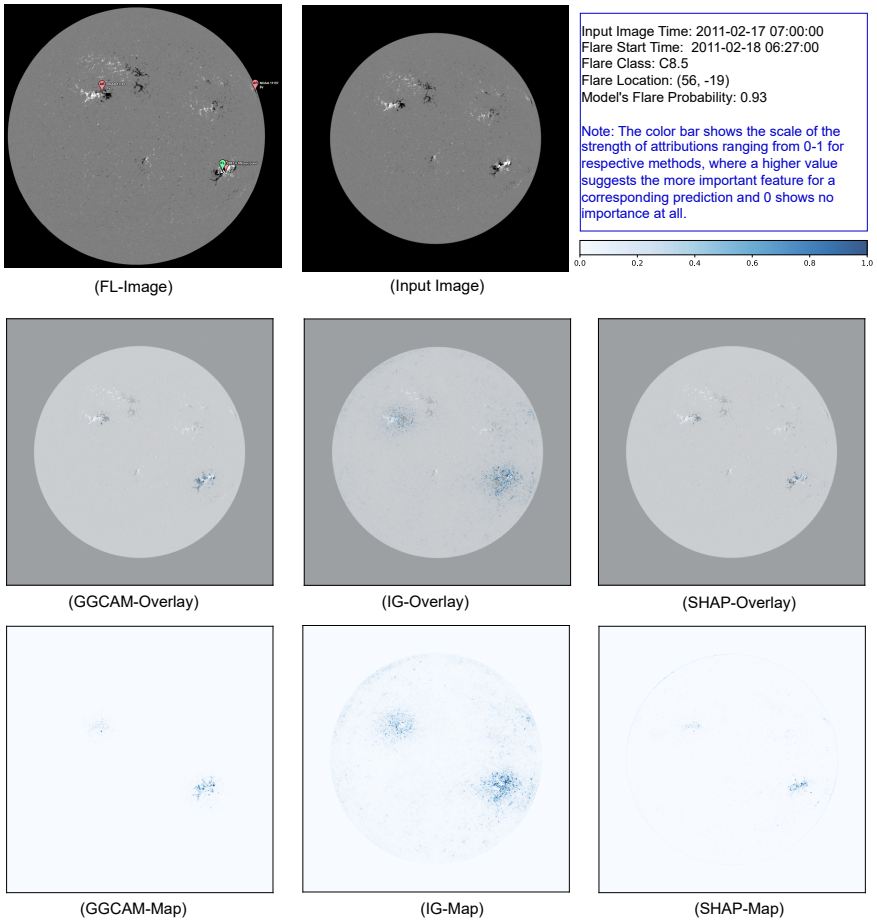

Fig. 7. A visual explanation of incorrectly predicted NF-class instance. (FL-Image): Annotated full-disk magnetogram at flare start time, showing flare location (green flag) and NOAA ARs (red flags). (Input Image): Actual magnetogram from the dataset. Overlays(GGCAM, IG, SHAP) depict the input image overlayed with attributions, and Maps(GGCAM, IG, SHAP)) showcase the attribution maps obtained from Guided Grad-CAM, Integrated Gradients, and Deep SHAP, respectively.

as soon as a region becomes visible, the pixels covering the AR on the east-limb are activated. 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 2013-10-29T21:42:00 UTC where we used an input image at 2013-10-29T03:00:00 UTC (∼19 hours prior to the flare event) shown in Fig. 6. We observed that the model focuses on specific ARs including the relatively smaller AR on the west limb, even though other ARs are present in the magnetogram

image. This shows that our models are capable of identifying the relevant AR even when there is a severe projection effect.

Similarly, to analyze a case of false positive, we present an example of a C8.5 flare observed on 2011-02-18 at 06:27:00 UTC, and to explain the result, we used an input magnetogram instance at 2014-02-17 07:00:00 UTC (∼23.5 hours prior to the event). We observed that the model's flaring probability for this particular instance is about 0.93. Therefore, we seek a visual explanation of this prediction using all three interpretation methods. Similar to the observations from our positive prediction, the visualization rendered using Guided Grad-CAM and Deep SHAP provides smoothed details and reveals that out of three ARs present in the magnetogram, only two of them are activated while the AR on the west-limb is not considered important for this prediction as shown in Fig. 7. Although an incorrect prediction, the visual explanation shows that the model's decision is based on an AR which is, in fact, responsible for the eventual C8.5 flare event. This incorrect prediction can be attributed to the interference of these bordering class flares, which is problematic for binary flare prediction models.

7 Conclusion and Future Work

In this paper, we employed three recent gradient-based attribution methods to interpret the predictions made by our binary flare prediction model based on VGG-16, which was trained to predict ≥M-class flares. We addressed the issue of flares occurring in near-limb regions of the Sun, which has been widely ignored, and our model demonstrated competent performance for such events. Additionally, we assessed the model's predictions with visual explanations, indicating that the decisions were primarily based on characteristics related to ARs in the magnetogram instance. Despite the model's enhanced ability, it still suffers from a high false positive rate due to high C-class flares. In an effort to address this problem, we plan to examine the unique features of each flare class to create a more effective method for segregating these classes based on background flux and generate a new set of labels that better handle border class flares. Moreover, our models currently only examine spatial patterns in our data, but we intend to broaden this work to include spatiotemporal models to improve performance.

Acknowledgements

This work 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.

Ethical Statement

Space weather forecasting research raises several ethical implications that must be considered. It is important to note that the data used for the full-disk deep learning model for solar flare prediction is publicly available as a courtesy of NASA/SDO and the AIA, EVE, and HMI science teams – and not subject to data privacy and security concerns. The use of SDO images for non-commercial purposes and public education and information efforts is strongly encouraged and requires no expressed authorization. However, it is still essential to consider the ethical implications associated with developing and using a full-disk deep learning model for solar flare prediction, particularly in terms of fairness, interpretability, and transparency. It is crucial to ensure that the model is developed and used ethically and responsibly to avoid any potential biases or negative impacts on individuals or communities. Moreover, post hoc analysis for full-disk deep learning models for solar flare prediction should avoid giving wrongful assumptions of causality and false trust. While these models may have robust and novel forecast skills, it is crucial to understand the scarcity of extreme solar events and the skill scores used to assess model performance. We note that these models are not perfect and have limitations that should be considered when interpreting their predictions. Therefore, it is important to use these models with caution and to consider multiple sources of information when making decisions, especially when in operations, related to space weather events. By being transparent about the limitations and uncertainties associated with these models, we can ensure that they are used ethically and responsibly to mitigate any potential harm to individuals or communities.

Furthermore, the impact of space weather events can range from minor disruptions to significant damage to critical infrastructure, such as power grids, communication systems, and navigation systems, with the potential to cause significant economic losses. Therefore, it is crucial to ensure public safety, particularly for astronauts and airline crew members, by providing information about potential dangers associated with space weather events. Finally, it is imperative to ensure that space weather forecasting research is used for peaceful purposes, i.e., early detection and in part avoiding vulnerabilities that may be caused by extreme space weather events.

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 Astrophysical Journal Supplement Series **254**(2), 23 (May 2021)
2. Ahmadzadeh, A., Hostetter, M., Aydin, B., Georgoulis, M.K., Kempton, D.J., Mahajan, S.S., Angryk, R.: Challenges with extreme class-imbalance and temporal coherence: A study on solar flare data. In: 2019 IEEE Intl. Conf. on Big Data (Big Data). IEEE (Dec 2019). https://doi.org/10.1109/bigdata47090.2019.9006505

3. 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 Astrophysical Journal **898**(2), 98 (Jul 2020)
4. Bobra, M.G., Couvidat, S.: Solar flare prediction using≤SDO≤/HMI VECTOR MAGNETIC FIELD DATA WITH a MACHINE-LEARNING ALGORITHM. The Astrophysical Journal **798**(2), 135 (Jan 2015). https://doi.org/10.1088/0004-637x/798/2/135
5. 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
6. Crown, M.D.: Validation of the NOAA space weather prediction center's solar flare forecasting look-up table and forecaster-issued probabilities. Space Weather **10**(6), n/a–n/a (Jun 2012). https://doi.org/10.1029/2011sw000760
7. Devos, A., Verbeeck, C., Robbrecht, E.: Verification of space weather forecasting at the regional warning center in belgium. Journal of Space Weather and Space Climate **4**, A29 (2014). https://doi.org/10.1051/swsc/2014025
8. 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). https://doi.org/10.1007/s11214-010-9701-8
9. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable AI methods - a brief overview. In: xxAI - Beyond Explainable AI, pp. 13–38. Springer International Publishing (2022). https://doi.org/10.1007/978-3-031-04083-2_2
10. Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jul 2017)
11. 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 Astrophysical Journal **856**(1), 7 (Mar 2018). https://doi.org/10.3847/1538-4357/aaae00
12. 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). https://doi.org/10.1109/bigdata50022.2020.9377906
13. Ji, A., Cai, X., Khasayeva, N., Georgoulis, M.K., Martens, P.C., Angryk, R.A., Aydin, B.: A systematic magnetic polarity inversion line data set from SDO/HMI magnetograms. The Astrophysical Journal Supplement Series **265**(1), 28 (Mar 2023). https://doi.org/10.3847/1538-4365/acb43a
14. Ji, A., Wen, J., Angryk, R., Aydin, B.: Solar flare forecasting with deep learning-based time series classifiers. In: 2022 26th International Conference on Pattern Recognition (ICPR). IEEE (Aug 2022). https://doi.org/10.1109/ICPR56361.2022.9956097
15. Korsós, M.B., Georgoulis, M.K., Gyenge, N., Bisoi, S.K., Yu, S., Poedts, S., Nelson, C.J., Liu, J., Yan, Y., Erdélyi, R.: Solar flare prediction using magnetic field diagnostics above the photosphere. The Astrophysical Journal **896**(2), 119 (Jun 2020). https://doi.org/10.3847/1538-4357/ab8fa2
16. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks (2014)

17. Kusano, K., Iju, T., Bamba, Y., Inoue, S.: A physics-based method that can predict imminent large solar flares. Science **369**(6503), 587–591 (Jul 2020). https://doi.org/10.1126/science.aaz2511
18. Lee, K., Moon, Y.J., Lee, J.Y., Lee, K.S., Na, H.: Solar flare occurrence rate and probability in terms of the sunspot classification supplemented with sunspot area and its changes. Solar Physics **281**(2), 639–650 (Sep 2012). https://doi.org/10.1007/s11207-012-0091-9
19. Leka, K., Barnes, G., Wagner, E.: The NWRA classification infrastructure: description and extension to the discriminant analysis flare forecasting system (DAFFS). Journal of Space Weather and Space Climate **8**, A25 (2018). https://doi.org/10.1051/swsc/2018004
20. Li, X., Zheng, Y., Wang, X., Wang, L.: Predicting solar flares using a novel deep convolutional neural network. The Astrophysical Journal **891**(1), 10 (Feb 2020). https://doi.org/10.3847/1538-4357/ab6d04
21. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1) (2021). https://doi.org/10.3390/e23010018
22. 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)
23. Muller, D., Fleck, B., Dimitoglou, G., Caplins, B., Amadigwe, D., Ortiz, J., Wamsler, B., Alexanderian, A., Hughitt, V., Ireland, J.: JHelioviewer: Visualizing large sets of solar images using JPEG 2000. Computing in Science & Engineering **11**(5), 38–47 (Sep 2009). https://doi.org/10.1109/mcse.2009.142
24. 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). https://doi.org/10.1109/msp.2022.3142719
25. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Ishii, M.: Deep flare net (DeFN) model for solar flare prediction. The Astrophysical Journal **858**(2), 113 (May 2018). https://doi.org/10.3847/1538-4357/aab9a7
26. 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 Astrophysical Journal **835**(2), 156 (jan 2017). https://doi.org/10.3847/1538-4357/835/2/156
27. Pandey, C., Angryk, R., 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
28. 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
29. 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
30. Park, E., Moon, Y.J., Shin, S., Yi, K., Lim, D., Lee, H., Shin, G.: Application of the deep convolutional neural network to the forecast of solar flare occurrence

using full-disk solar magnetograms. The Astrophysical Journal **869**(2), 91 (Dec 2018)
31. 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
32. 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
33. Shapley, L.: A Value for N-Person Games. RAND Corporation (1952)
34. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences (2019)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
36. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net (2014). https://doi.org/10.48550/ARXIV.1412.6806
37. 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
38. 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 Astrophysical Journal **931**(2), 163 (Jun 2022). https://doi.org/10.3847/1538-4357/ac64a6, https://doi.org/10.3847/1538-4357/ac64a6
39. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017). https://doi.org/10.48550/ARXIV.1703.01365
40. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2015)
41. Toriumi, S., Wang, H.: Flare-productive active regions. Living Reviews in Solar Physics **16**(1) (May 2019). https://doi.org/10.1007/s41116-019-0019-7
42. 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

43. 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 Astrophysical Journal **910**(1), 8 (Mar 2021). https://doi.org/10.3847/1538-4357/abdebe