Towards Interpretable Solar Flare Prediction with Attention-based Deep Neural Networks

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

***Abstract*—Solar flare prediction is a central problem in space weather forecasting and recent developments in machine learning and deep learning accelerated the adoption of complex models for data-driven solar flare forecasting. In this work, we developed an attention-based deep learning model as an improvement over the standard convolutional neural network (CNN) pipeline to perform full-disk binary flare predictions for the occurrence of $\geq$M1.0-class flares within the next 24 hours. For this task, we collected compressed images created from full-disk line-of-sight (LoS) magnetograms. We used data-augmented oversampling to address the class imbalance issue and used true skill statistic (TSS) and Heidke skill score (HSS) as the evaluation metrics. Furthermore, we interpreted our model by overlaying attention maps on input magnetograms and visualized the important regions focused on by the model that led to the eventual decision. The significant findings of this study are: (i) We successfully implemented an attention-based full-disk flare predictor ready for operational forecasting where the candidate model achieves an average TSS=0.54$\pm$0.03 and HSS=0.37$\pm$0.07. (ii) we demonstrated that our full-disk model can learn conspicuous features corresponding to active regions from full-disk magnetogram images, and (iii) our experimental evaluation suggests that our model can predict near-limb flares with adept skill and the predictions are based on relevant active regions (ARs) or AR characteristics from full-disk magnetograms.**

***Index Terms*—space weather, solar flares, deep neural networks, attention, and interpretability.**

I. Introduction

Solar flares are relatively short-lasting events, manifested as the sudden release of huge amounts of energy with significant increases in extreme ultraviolet (EUV) and X-ray fluxes, and are one of the central phenomena in space weather forecasting. They are detected by the X-ray Sensors (XRS) instrument onboard Geostationary Operational Environmental Satellite (GOES) [1] and classified according to their peak X-ray flux level, measured in watts per square meter (Wm^{-2}) into the following five categories by the National Oceanic and Atmospheric Administration (NOAA): X ($\geq 10^{-4}Wm^{-2}$), M ($\geq 10^{-5}$ and $< 10^{-4}Wm^{-2}$), C ($\geq 10^{-6}$ and $< 10^{-5}Wm^{-2}$), B ($\geq 10^{-7}$ and $< 10^{-6}Wm^{-2}$), and A ($\geq 10^{-8}$ and $< 10^{-7}Wm^{-2}$) [2]. In solar flare forecasting, M- and X-class flares are large and relatively scarce events and are usually considered to be the class of interest as they are more likely to have a near-Earth impact that can affect both space-based systems (e.g., satellite communication systems) and ground-based infrastructures (e.g., electricity supply chain and airline industry) and even pose radiation hazards to astronauts in space. Therefore, it is essential to have a precise and reliable approach for predicting solar flares to mitigate the associated life risks and infrastructural damages.

Active regions (ARs) are the areas on the Sun (visually indicated by scattered red flags in full-disk magnetogram image, shown in Fig. 1) with disturbed magnetic field and are considered to be the initiators of various solar activities such as coronal mass ejections (CMEs), solar energetic particle (SEP) events, and solar flares [3]. The majority of the approaches for flare prediction primarily target these ARs as regions of interest and generate predictions for each AR individually. The magnetic field measurements, which are the dominant feature employed by the AR-based forecasting techniques, are susceptible to severe projection effects as ARs get closer to limbs to the degree that after $\pm 60^\circ$ the magnetic field readings are distorted [4]. Therefore, the aggregated flare occurrence probability (for the whole disk), in fact, is restricted by the capabilities of AR-based models. This is because the input data is restricted to ARs located in an area within $\pm 30^\circ$ (e.g., [5]) to $\pm 70^\circ$ (e.g., [6]) from the center due to severe projection effects [7]. As AR-based models include data up to $\pm 70^\circ$, in the context of this paper, this upper limit ($\pm 70^\circ$) is used as a boundary for central location (within $\pm 70^\circ$) and near-limb regions (beyond $\pm 70^\circ$) as shown in Fig. 1.

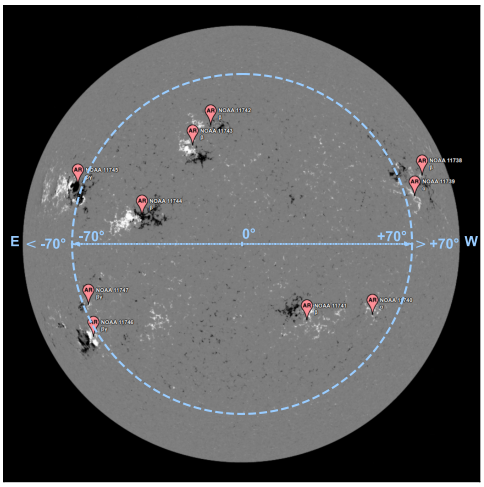

Fig. 1. An annotated full-disk magnetogram image as observed on 2013-05-13 at 02:00:00 UTC, showing the approximate central location (within $\pm 70^\circ$) and near-limb (beyond $\pm 70^\circ$ to $\pm 90^\circ$) region with all the visible active regions present at the noted timestamp, indicated by the red flags. Note that the directions East (E) and West (W) are reversed in solar coordinates.

Furthermore, to issue a full-disk forecast using an AR-based model, the usual approach involves aggregating the flare probabilities from each AR by applying a heuristic function, as outlined in [8]. This aggregated result estimates the probability of at least one AR experiencing a flare event, assuming that the occurrence of flares in different ARs is conditionally independent and assigning equal weights to each AR during full-disk aggregation. This uniform weighting approach may not accurately capture the true impact of each AR on the probability of predicting full-disk flares [9]. It is essential to note that the specific weights for these ARs are generally unknown, and there are no established methods for precisely determining these weights. While AR-based models are limited to central locations and require a heuristic to aggregate and issue comprehensive forecasts, full-disk models use complete, often compressed, magnetograms corresponding to the entire solar disk. These magnetograms are used for shape-based parameters such as size, directionality, sunspot borders [10], and polarity inversion lines [11]. Although projection effects still prevail in the original magnetogram rasters, deep-learning models can learn from the compressed full-disk images as observed in [12]–[14] and issue the flare forecast for the entire solar disk. Therefore, a full-disk model is appropriate to complement the AR-based counterparts as these models can predict the flares that appear on the near-limb regions of the Sun and add a crucial element to the operational systems.

Deep learning-based approaches have significantly improved results in generic image classification tasks; however, these models are not easily interpretable due to the complex modeling that obscures the rationale behind the model's decision. Understanding the decision-making process is critical for operational flare forecasting systems. 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) (e.g., [15]), meaning they focus on the analysis of trained models and do not contribute to the model's parameters while training. In this work, we primarily focus on developing a convolutional neural network (CNN) based full-disk model with trainable attention modules that can amplify the relevant features and suppress the misleading ones while predicting $\geq$M1.0-class solar flares as well as evaluating and explaining our model's performance by visualizing the attention maps overlaying on the input magnetograms to understand which regions on the magnetogram were considered relevant for the corresponding decision. To validate and compare our results, we train a baseline model with the same architecture as our attention model, which however, follows the standard CNN pipeline where a global image descriptor for an input image is obtained by flattening the activations of the last convolutional layer.

By integrating attention modules into the standard CNN pipeline, we attain two significant advantages: enhanced model performance and the ability to gain insight into the decision-making process. This integration not only improves the predictive abilities but also provides an interpretable model that reveals the significant features influencing the model's decisions. The architecture combines the CNN pipeline with trainable attention modules as mentioned in [16]. Both of our model's architectures are based on the general CNN pipeline; details are described later in Sec. IV. The novel contributions of this paper are as follows: (i) We introduce a novel approach of a light-weight attention-based model that improves the predictive performance of traditional CNNs for full-disk solar flare prediction (ii) We utilize the attention maps from the model to understand the model's rationale behind prediction decision and show that the model's decisions are linked to relevant ARs (iii) We show that our models can tackle the prediction of flares appearing on near-limb regions of the Sun.

The remainder of this paper is organized as follows: In Sec. II, we outline the various approaches used in solar flare prediction with contemporary work using deep learning. In Sec. III, we explain our data preparation and class-wise distribution for binary prediction mode. In IV we present a detailed description of our flare prediction model. In Sec. V, we present our experimental design and evaluations. In Sec. VI we present case-based qualitative interpretations of attention maps, and, lastly, in Sec. VII, we provide our concluding remarks with avenues for future work.

II. Related Work

Solar flare prediction currently, to the best of our knowledge, relies on four major strategies: (i) empirical human prediction (e.g., [17], [18]), which involves manual monitoring and analysis of solar activity using various instruments and techniques, to obtain real-time information about changes in the Sun's magnetic field and surface features, which are often precursors to flare activity; (ii) physics-based numerical simulations (e.g., [19], [20]), which involves a detailed understanding of the Sun's magnetic field and the processes that drive flare activity and running simulations models to predict the occurrence of flares; (iii) statistical prediction (e.g., [21], [22]), which involves studying the historical behavior of flares to predict their likelihood in the future using statistical analysis and is closely related to (iv) machine learning and deep learning approaches (e.g., [5], [6], [8], [23]–[30]), which involves training algorithms with vast amount of historical data and creating data-driven models that detects subtle patterns associated with flares in solar activity and make predictions.

The rapid advancements in deep learning techniques have significantly accelerated research in the field of solar flare prediction. A CNN-based flare forecasting model trained with solar AR patches extracted from line-of-sight (LoS) magnetograms within $\pm 30^\circ$ of the central meridian to predict $\geq$C-, $\geq$M-, and $\geq$X-class flares was presented in [5]. Similarly, [26] use a CNN-based model to issue binary class predictions for both $\geq$C- and $\geq$M-class flares within 24 hours using Space-Weather Helioseismic and Magnetic Imager Active Region Patches (SHARP) data [31] extracted from solar magnetograms using AR patches located within $\pm 45^\circ$ of the central meridian. Both of these models are limited to a small portion of the observable disk in central locations ($\pm 30^\circ$ and $\pm 45^\circ$) and thus have limited operational capability.

Moreover, in our previous studies [27], [28], we presented deep learning-based full-disk flare prediction models. These models were trained using smaller datasets and these proof-of-concept models served as initial investigations into their potential as a supplementary component for operational forecasting systems. More recently, we presented explainable full-disk flare prediction models [12], [13], utilizing attribution methods to comprehend the models' effectiveness for near-limb flare events. We observed that the deep learning-based full-disk models are capable of identifying relevant areas in a full-disk magnetogram, which eventually translates into the model's prediction. However, these models utilized a post-hoc approach for model explanation, which does not contribute to further improving the model's performance.

In recent years, attention-based models, particularly Vision Transformers (ViTs) [32], have emerged as powerful contenders for image classification tasks, achieving competent results on large-scale datasets. ViTs leverage self-attention mechanisms to effectively capture long-range dependencies in images, enabling them to excel in complex visual recognition tasks. While ViTs offer state-of-the-art performance, they often come with a large number (86 to 632 million) of trainable parameters, making them resource-intensive and less practical for scenarios with limited computational resources or small-sized datasets. To address this issue, for our specific use case with a small dataset, we are exploring alternative models that strike a balance between accuracy and efficiency. By incorporating attention blocks into a standard CNN pipeline, we obtain a much lighter model, consisting of ∼7.5 million parameters. This approach allows for computationally efficient near-real-time predictions with relatively less resource demand on deployment infrastructure while ensuring competent performance for solar flare prediction compared to our prior work [13], [14] with customized AlexNet-based [33] full-disk model, with ∼57.25 million parameters and fine-tuned VGG16 [34] full-disk model in [12] with ∼134 million parameters.

III. Data

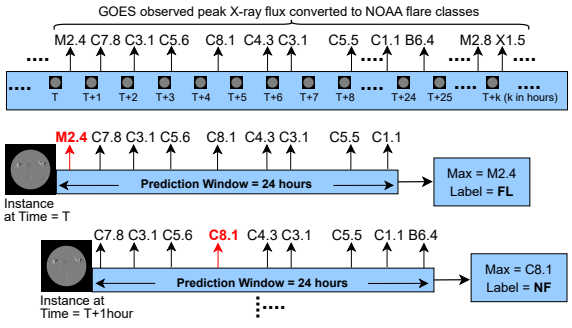

Fig. 2. A visual representation of the data labeling process using hourly observations of full-disk LoS magnetograms and a prediction window of 24 hours considered to label the magnetograms. Here, 'FL' and 'NF' indicate 'Flare' and 'No Flare' for binary prediction mode (≥M1.0-class flares).

We use full-disk line-of-sight (LoS) magnetogram images obtained from the Helioseismic and Magnetic Imager (HMI) [35] instrument onboard Solar Dynamics Observatory (SDO) [36] publicly available from Helioviewer [37]. We collected hourly instances of magnetogram images at [00:00, 01:00,...,23:00] each day from December 2010 to December 2018. We labeled the magnetogram images for binary prediction mode (≥M1.0-class flares) based on the peak X-ray flux converted to NOAA flare classes with a prediction window of the next 24 hours. To elaborate, if the maximum of GOES observed peak X-ray flux of a flare is weaker than M1.0, the corresponding magnetogram instances are labeled as "No Flare" (NF: <M1.0), and larger ones are labeled as "Flare" (FL: ≥M1.0) as shown in Fig .2.

TABLE I
THE TOTAL NUMBER OF HOURLY SAMPLED MAGNETOGRAM IMAGES PER FLARE CLASS DISTRIBUTED INTO FOUR TRI-MONTHLY PARTITIONS.

Binary Class	Partition-1	Partition-2	Partition-3	Partition-4	Total
NF (<M1.0)	12,454	13,855	14,308	14,032	54,649
FL (≥M1.0)	2,334	1,612	2,364	2,690	9,000
FL:NF	∼1:5	∼1:9	∼1:6	∼1:5	∼1:6

Our dataset includes a total of 63,649 full-disk LoS magnetogram images, where 54,649 instances belong to the NF-class and 9,000 instances (8,120 instances of M-class and 880 instances of X-class flares) to the FL-class [1]. We finally create a non-chronological split of our data into four temporally non-overlapping tri-monthly partitions introduced in [27] for our 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 January to March, Partition-2 contains data from April to June, Partition-3 contains data from July to September, and finally, Partition-4 contains data from October to December as shown in Table. I. Because ≥M1.0-class flares are scarce, the data distribution exhibits a significant imbalance, with the highest imbalance occurring in Partition-2 (FL:NF ∼1:9). Overall, the imbalance ratio stands at ∼1:6 for FL to NF class.

IV. Model

In this work, we develop two deep learning models: (i) standard CNN model as a baseline (denoted as *M1*), and (ii) attention-based model (denoted as *M2*) proposed in [16] to perform and compare in the task of solar flare prediction. The M1 model shown in Fig. 3 follows an intuition of standard CNN architecture where a global image descriptor (g) is derived from the input image from the activations of the last convolutional layer and passed through a fully connected layer to obtain class prediction probabilities. On the other hand, the attention-based full-disk model (M2) encourages the filters earlier in the CNN pipeline to learn similar mappings compatible with the one that produces a global image descriptor in the original architecture. Furthermore, it focuses on identifying salient image regions and amplifying their influence while

[1]The current total count of 63,649 magnetogram observations in our dataset is lower than it should be for the period of December 2010 to December 2018. This is due to the unavailability of some instances from Helioviewer.

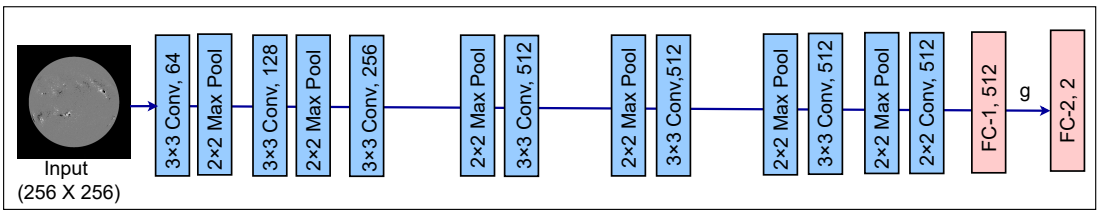

Fig. 3. The architecture of our baseline model (M1). Note: Each convolutional layer (except the last one) is followed by a batch normalization layer.

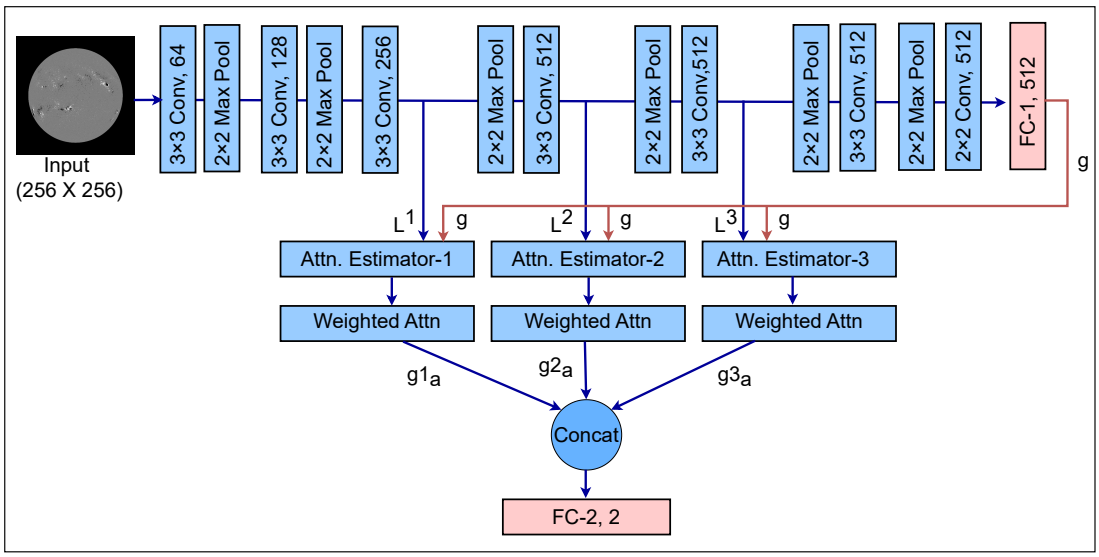

Fig. 4. The architecture of our attention-based flare prediction model (M2). The model has three trainable attention modules integrated after the third, fourth, and fifth convolution blocks before the max-pool layer. Note: Each convolutional layer (except the last one) is followed by a batch normalization layer.

suppressing the irrelevant and potentially spurious information in other regions during training and thus utilizing a trainable attention estimator by integrating it into the standard CNN pipeline. The architecture of our attention-based model is shown in Fig. 4. The architecture of the attention model proposed in [16] integrates the trainable attention modules in a modified VGG-16 [34] architecture. We use a simpler VGG-like architecture with a reduced number of convolutional layers, which also reduces the number of parameters. Our first convolutional layer accepts a 1-channel input magnetogram image resized to 256×256. Each convolutional layer (except the last one) is followed by a batch normalization layer before max pooling. The final convolutional layer outputs feature maps of size 512×1×1 that squeezed into a fully connected layer (FC-1) with a 512-dimensional vector, which is the global representation (g) of the input image.

The M2 model follows the same architecture as in M1, except it has three trainable attention modules integrated after the third, fourth, and fifth convolution blocks before the max-pool layer. The similarity between the architectures is intentional to demonstrate the impact of the attention estimators on model performance. Similarly, integrating attention modules in the middle of the network is also a deliberate design choice. As the early layers in CNN primarily focus on low-level features [38], we position the attention modules further into the pipeline to capture higher-level features. However, there is a tradeoff involved, as pushing attention to the last layers is hindered by significantly reduced spatial resolution in the feature maps. Consequently, placing attention modules in the middle strikes a balance, making it a more suitable and pragmatic approach.

In the M2 model, outputs from the convolutional blocks (denoted as L^s) are passed to the attention estimators. In other words, L^s is a set of feature vectors:

$$L^s = \{l_1^s, l_2^s, ..., l_n^s\}$$

extracted at a given convolutional layer to serve as input to the s_{th} attention estimator, and l_i^s is the vector of output activations at i^{th} of total n spatial locations in the layer. g represents a global feature vector obtained by flattening the feature maps at the first fully connected layer, located at the end of the convolution blocks (referred to as FC-1 in Fig.4).

The attention mechanism aims to compute a compatibility score, denoted as $C(L^s, g)$, utilizing the local features (L^s) and global feature representations (g), and replaces the final feature vector with a set of attention-weighted local features. As the compatibility scores C and L^s are required to have the same dimension, the dimension matching is performed by a linear mapping of vectors l_i^s to the dimension of g. Then, the compatibility function $C(L^s, g) = \{c_1^s, c_2^s, ..., c_n^s\}$ is a set for each vector l_i^s, which is computed as an addition operation (additive attention) as follows:

$$c_1^s = (l_i^s, g), \text{ for } i \in \{1, 2, ..., n\}.$$

The computed compatibility scores are then normalized using a softmax operation and represented as:

$$A^s = \{a_1^s, a_2^s, ..., a_n^s\}.$$

The normalized compatibility scores are then used to compute an element-wise weighted average, which results in a vector:

$$g_a^s = \sum_{i=1}^{n} a_i^s . l_i^s$$

for each attention layer, s. Finally, the individual g_a^s vectors of size 512 are concatenated to get a new attention-based global representation to perform the binary classification in the (second) fully connected layer (FC-2). This approach allows the activations from earlier layers to influence and contribute to the final global feature representation, thereby enhancing the model's ability to capture relevant spatial information.

V. Experimental Evaluation

A. Experimental Settings

We trained both of our models (M1 and M2) with stochastic gradient descent (SGD) as an optimizer and cross-entropy as the objective function. Both models are initialized using Kaiming initialization from a uniform distribution [39], and then we use a dynamic learning rate (initialized at 0.001 and reduced by half every 3 epochs) to further train the model to 40 epochs with a batch size of 128. We regularized our models with a weight decay parameter tuned at 0.5 to prevent overfitting. As mentioned earlier in Sec. III, we are dealing with an imbalanced dataset. Therefore, we address the class imbalance problem through data augmentation and oversampling exclusively for the training set while maintaining the imbalanced nature of the test set for realistic evaluation. Firstly, we use three augmentation techniques: vertical flipping, horizontal flipping, and +5° to -5° rotations on minority class (FL-class) which decreases the imbalance from 1:6 to approximately 2:3. Finally, we randomly oversampled the minority class to match the instances of NF-class resulting in a balanced dataset. We prefer augmentation and oversampling over undersampling as the flare prediction models trained with undersampled data are shown to lead to inferior performance [40] (usually transpiring as one-sided predictions). We employed a 4-fold cross-validation schema for validating our models, using the tri-monthly partitions (described in Sec. III), where we applied three partitions for training the model and one for testing.

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 this paper, the "FL-class" is considered as the positive outcome, while the "NF-class" is 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. 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. For a balanced test dataset, these metrics are equivalent [40]. 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.

B. Evaluation

TABLE II
Average performance of our models in terms of two skill scores (TSS and HSS) evaluated on the test set for the 4-fold cross-validation experiment.

Models	TSS	HSS
M1	0.35±0.13	0.30±0.09
Pandey et al. [12]	∼0.51	∼0.35
Pandey et al. [13]	0.51±0.05	0.38±0.08
M2	**0.54±0.03**	**0.37±0.07**

We perform a 4-fold cross-validation using the tri-monthly separated dataset for evaluating our models. With the baseline model (M1) we obtain on an average TSS∼0.35±0.13 and HSS∼0.30±0.09. The M1 model following the standard CNN pipeline has fluctuations across folds and hence a high margin of error on skill scores is represented by the standard deviation. Model M2 improves over the performance of model M1 by ∼20% and ∼7% in terms of TSS and HSS respectively. Furthermore, it improves on the performance of [12], [13] by ∼3% in terms of TSS and shows comparable results in terms of HSS and is more robust as indicated by the deviations across the folds as shown in Table II. Moreover, the performance of model M2 becomes even more noteworthy when considering its parameter efficiency. With only ∼7.5 million parameters, it outperforms [13] an AlexNet-based model and [12] a VGG16-based model with a much higher parameter count of ∼57.25 and ∼134 million respectively, showcasing the effectiveness of attention mechanisms in achieving superior results while maintaining a significantly leaner model architecture. This highlights the potential of this approach to provide both performance gains and resource optimization. The findings of this study emphasize the significance of optimizing attention configurations to enhance model performance, taking into account both parameter complexities and the strategic combination of attention patterns for effective pattern recognition.

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 III.

TABLE III
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°		
Models	Flare-Class	TP	FN	Recall	TP	FN	Recall
M1	X-Class	467	201	0.70	100	112	0.47
	M-Class	3153	2677	0.54	878	1412	0.38
	Total (X&M)	3620	2878	0.62	978	1524	0.43
M2	X-Class	636	32	**0.95**	164	48	**0.77**
	M-Class	4850	980	**0.83**	1161	1129	**0.51**
	Total (X&M)	5486	1012	**0.89**	1325	1177	**0.64**

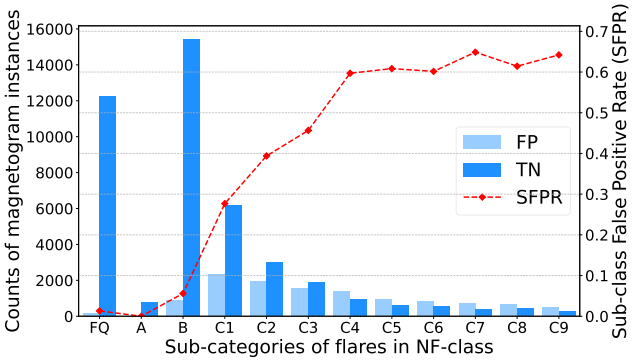

Fig. 5. A bar-line plot showing the true negatives (TN), false positives (FP), and false positive rate for sub-classes in NF-class (SFPR) obtained from model M2. The results are aggregated from validation sets of 4-fold experiments.

We observe that the attention-based model (M2) shows significantly better results compared to the baseline (M1). The M2 model made correct predictions for ∼95% of the X-class flares and ∼83% of the M-class flares in central locations. Similarly, it shows a compelling performance for flares appearing on near-limb locations of the Sun, where ∼77% of the X-class and ∼51% of the M-class flares are predicted correctly. This is important because, to our knowledge, the prediction of near-limb flares is often overlooked, although vital for predicting Earth-impacting space weather events. More false negatives in M-class are expected because of the model's inability to distinguish bordering class (C4+ to C9.9) flares from ≥M1.0-class flares as shown in Fig. 5. We observed an upward trend in the false positive rate for sub-classes (SFPR) within C-class flares when compared to other sub-classes, such as Flare-Quiet (FQ), A-class, and B-class flares. More specifically we note that the count of false positives (FP) surpasses that of true negatives (TN) for flare classes ranging from ≥C4 to ≤C9. The prevalence of FP in ≥C4-class flares suggests a need for improved predictive capabilities between border classes.

Overall, we observed that our model predicted ∼89% of the flares in central locations and ∼64% of the flares in near-limb locations. Furthermore, class-wise analysis shows that ∼91% and ∼74% of the X-class and M-class flares, respectively, are predicted correctly by our models. To reproduce this work, the source code is available in our open-source repository [41].

VI. DISCUSSION

In this section, we visualize the attention maps learned by the M2 model to qualitatively analyze and understand regions in input magnetogram images that are considered relevant. We applied three attention layers in our model M2, where attention maps $(L1, L2, L3)$ has a spatial dimension $(\frac{1}{4}, \frac{1}{8}, \frac{1}{16})th$ of the input size respectively. To visualize the relevant features learned by the models using attention layers, we upscale these maps to the size of the magnetogram image using bilinear interpolation and overlay the maps on top of the original image. We present the attention maps from the Attention Estimator-2 because the first attention layer focuses on lower-level features, which are scattered and do not provide a globally detailed explanation. On the other hand, the Attention Estimator-3 focuses on higher-level features, and due to the high reduction in spatial dimension ($\frac{1}{16}$ of the original input), upscaling through interpolation results in a spatial resolution that is insufficient for generating interpretable activation maps.

As the primary focus of this study is to understand the capability of full-disk models on the near-limb flares, we showcase a near-limb (East) X3.2-class flare observed on 2013-05-14T00:00:00 UTC. Note that East and West are reversed in solar coordinates. The location of the flare is shown by a green flag in Fig. 6 (a)(i), along with the ARs (red flags). For this case-based qualitative analysis, we use an input image at 2013-05-13T06:00:00 UTC (∼18 hours prior to the flare event), shown in Fig. 6 (a)(ii) and in Fig. 6 (a)(iii), we show the overlaid attention map, which pinpoints important regions in the input image where specific ARs are activated as relevant features, suppressing a large section of the full-disk magnetogram disk although there are 10 ARs (red flags). More specifically, the model focuses on the same AR that is responsible for initiating a flare 18 hours later. Similarly, we analyze another case of correctly predicted near-limb (West) X1.0-class flare observed on 2013-11-19T10:14:00 UTC shown in Fig. 6 (b)(i). For this, we used an input image at 2013-11-18T17:00:00 UTC (∼17 hours prior to the flare event) shown in Fig. 6 (b)(ii). We again observed that the model focuses on the relevant AR even though other, relatively large ARs are present in the magnetogram image as shown in Fig. 6 (b)(iii).

Furthermore, we provide an example to analyze a case of false positives as well. For this, we use an example of a C7.9 flare observed on 2014-02-03T00:12:43 UTC shown in Fig .6 (c)(i), and to explain the result, we used an input magnetogram instance at 2014-02-02T23:00:00 UTC (∼14 hours prior to the event) shown in Fig .6 (c)(ii). For the given time, there are 7 ARs indicated by the red flags, however, on interpreting this prediction with attention maps shown in Fig .6 (c)(iii), we observed that the model considers only one region as a relevant feature for the corresponding prediction, which is indeed the location of the C7.9 flare. This incorrect prediction can be attributed to interference caused by bordering C-class flares as shown earlier in Fig. 5, where we noted that among the 25,150 C-class flares observed, ∼43% (10,935) resulted in incorrect predictions, constituting ∼91% of the total false positives.

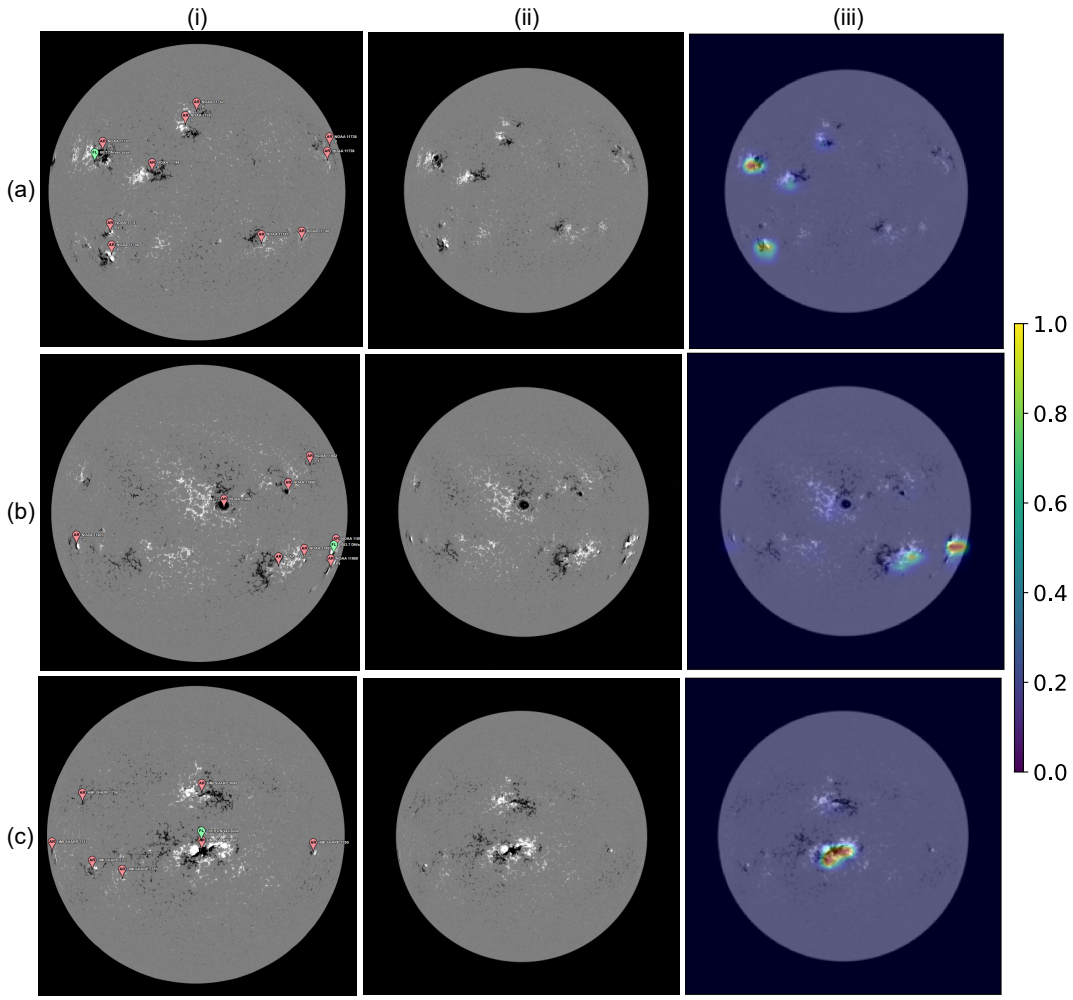

Fig. 6. A figure-grid of case-based visual interpretation for three different instances using attention maps, each represented in a separate row indexed as (a), (b), and (c). Row (a) shows correctly predicted near-limb (East) FL-class, (b) shows correctly predicted near-limb (West) FL-class, and (c) shows incorrectly predicted NF-class instances. In each row: column (i) displays an annotated full-disk magnetogram image at the onset of the flare, with green flags indicating the flare's location and red flags representing all ARs present in the magnetogram. Column (ii) shows an actual magnetogram image used in our dataset to train the model. Finally, column (iii) depicts a visualization created by overlaying the attention maps obtained from Attention Estimator-2. The color bar shows the scale of normalized attention map values ranging from 0-1, where a higher value suggests important features for a corresponding prediction.

VII. Conclusion and Future Work

In this work, we presented an attention-based full-disk model to predict $\geq$M1.0-class flares in binary mode and compared the performance with standard CNN-based models. We observed that the trainable attention modules play a crucial role in directing the model to focus on pertinent features associated with ARs while suppressing irrelevant features in a magnetogram during training, resulting in an enhancement of model performance. Furthermore, we demonstrated, both quantitatively through recall scores and qualitatively by overlaying attention maps on input magnetogram images, that our model effectively identifies and localizes relevant AR locations, which are more likely to initiate a flare. This prediction capability extends to near-limb regions, making it crucial for operational systems. As an extension, we plan to include the temporal aspects in our dataset and create a spatiotemporal model to capture the evolution of solar activity leading to solar flares. Furthermore, we plan to extend this work by developing an automated way of analyzing the interpretation results to identify the main causes of incorrect predictions.

Acknowledgments

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. The data used in this study is a courtesy of NASA/SDO and the AIA, EVE, and HMI science teams, and the NOAA National Geophysical Data Center (NGDC).

REFERENCES

[1] P. C. Chamberlin, T. N. Woods, F. G. Eparvier, and A. R. Jones, "Next generation x-ray sensor (XRS) for the NOAA GOES-r satellite series," in *SPIE Proceedings*, S. Fineschi and J. A. Fennelly, Eds. SPIE, Aug. 2009.

[2] 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.

[3] Z. Hamidi, M. Noh, W. Toni, and N. Shariff, "An analysis of active region as a trigger of solar flares," *Journal of Physics: Conference Series*, vol. 1298, no. 1, p. 012017, Aug. 2019.

[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 ApJ*, 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 ApJ*, 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 *International Conference on 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] 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.

[9] C. Pandey, M. K. Georgoulis, B. Aydin, R. A. Angryk, and A. Ji, "Exploring heuristics in full-disk aggregation from individual active region prediction of solar flares," in *44th COSPAR Scientific Assembly. Held 16-24 July*, vol. 44, 2022, p. 3457.

[10] P. S. McIntosh, "The classification of sunspot groups," *Solar Physics*, vol. 125, pp. 251–267, 1990.

[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 ApJ Supplement Series*, vol. 265, no. 1, p. 28, 2023.

[12] 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

[13] 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

[14] C. Pandey, A. Ji, T. Nandakumar, R. A. Angryk, and B. Aydin, "Exploring deep learning for full-disk solar flare prediction with empirical insights from guided grad-cam explanations," 2023. [Online]. Available: https://arxiv.org/abs/2308.15712

[15] 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.

[16] S. Jetley, N. A. Lord, N. Lee, and P. H. S. Torr, "Learn to pay attention," 2018. [Online]. Available: https://arxiv.org/abs/1804.02391

[17] 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.

[18] 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.

[19] 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.

[20] 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 ApJ*, vol. 896, no. 2, p. 119, Jun. 2020.

[21] 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.

[22] 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.

[23] M. G. Bobra and S. Couvidat, "Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm," *The ApJ*, vol. 798, no. 2, p. 135, Jan. 2015.

[24] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, and M. Ishii, "Deep flare net (DeFN) model for solar flare prediction," *The ApJ*, vol. 858, no. 2, p. 113, May 2018.

[25] N. Nishizuka, Y. Kubo, K. Sugiura, M. Den, and M. Ishii, "Operational solar flare prediction model using deep flare net," *Earth, Planets and Space*, vol. 73, no. 1, Mar. 2021.

[26] X. Li, Y. Zheng, X. Wang, and L. Wang, "Predicting solar flares using a novel deep convolutional neural network," *The ApJ*, vol. 891, no. 1, p. 10, Feb. 2020.

[27] 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.

[28] 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.

[29] K. Whitman, R. Egeland, I. G. Richardson, and et al., "Review of solar energetic particle models," *Advances in Space Research*, Aug. 2022.

[30] 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.

[31] 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.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020.

[33] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014.

[35] 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.

[36] W. 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.

[37] D. Muller, B. Fleck, G. Dimitoglou, B. Caplins, D. Amadigwe, 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.

[38] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Mar. 2021.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," 2015.

[40] A. Ahmadzadeh, B. Aydin, M. Georgoulis, D. Kempton, S. Mahajan, and R. Angryk, "How to train your flare prediction model: Revisiting robust sampling of rare events," *The ApJ Supplement Series*, vol. 254, no. 2, p. 23, May 2021.

[41] DMLab, "Source Code." [Online]. Available: https://bitbucket.org/gsudmlab/fulldiskattention/src/main/