Unveiling the Potential of Deep Learning Models for Solar Flare Prediction in Near-Limb Regions

Abstract—**In recent years, the development of complex models for data-driven solar flare prediction has been accelerated by advancements in machine learning and deep learning utilizing a variety of approaches and data products, while most studies only address and assess the models' efficacy in central locations (within ±70° in longitude of the solar disk). This study aims to evaluate the performance of deep learning models in predicting ≥M-class solar flares with a prediction window of 24 hours, using hourly sampled full-disk line-of-sight (LoS) magnetogram images, particularly focusing on the often overlooked flare events corresponding to the near-limb regions (beyond ±70° in longitude of the full-disk) that constitute ∼40% of the total area of the entire solar disk. We performed our spatial and temporal analytical evaluations using three well-known deep learning architectures–AlexNet, VGG16, and ResNet34 using transfer learning. Furthermore, we compare and evaluate the overall performance of our models using true skill statistics (TSS) and Heidke skill score (HSS) and compute recall scores to understand the prediction sensitivity in central and near-limb regions for both X- and M-class flares. The following points summarize the key findings of our study: (1) The highest overall performance was observed with the AlexNet-based model, which achieved an average TSS of ∼0.53 and an HSS of ∼0.37; (2) Further, a meticulous spatial analysis of recall scores disclosed that for the near-limb events, the VGG16- and ResNet34-based models exhibited superior prediction sensitivity. The best results, however, were seen with the ResNet34-based model for the near-limb flares, where the average recall was roughly 0.59 (the recall for X- and M-class was 0.81 and 0.56 respectively) and (3) Our research findings demonstrate that our models are capable of discerning complex spatial patterns from full-disk magnetograms and exhibit skill in predicting solar flares, even in the vicinity of near-limb regions. This ability holds substantial importance for operational forecasting systems.**

Index Terms—**deep learning, solar flares, near-limb prediction**

I. Introduction

Solar flares are temporary occurrences on the Sun, considered to be the central phenomena in space weather forecasting, manifested as the sudden large eruption of electromagnetic radiation on the outermost atmosphere of the Sun. They are classified according to their peak X-ray flux level into the following five categories by National Oceanic and Atmospheric Administration (NOAA): X $(> 10^{-4}Wm^{-2})$, M $(> 10^{-5}Wm^{-2})$, C $(> 10^{-6}Wm^{-2})$, B $(> 10^{-7}Wm^{-2})$, and A $(> 10^{-8}Wm^{-2})$ [1]. M- and X-class solar flares are relatively scarce events and significantly more powerful than the other flare classes and, therefore, the class of interest that gathers the attention of researchers. These flares may potentially disrupt the electricity supply chain, airline industry, and satellite communications, and pose radiation hazards to astronauts in space [2].

Active regions (ARs) on the Sun are areas, visually indicated by scattered red flags in full-disk magnetogram image shown in Fig. 1, where the Sun's magnetic field is disturbed, and they spawn various types of solar activity such as solar flare, coronal mass ejection (CME), and solar energetic particle (SEP) events. Most operational flare forecasts [3] target these regions of interest and issue predictions for individual ARs, which are the main initiators of space weather events. To issue a full-disk forecast with an AR-based model, the output flare probabilities for each active region are usually aggregated using a heuristic function as mentioned in [4]. The heuristic function used to aggregate the final forecast operates under the assumption of conditional independence among active regions and that all active regions contribute equally to the aggregate forecast. This uniform weighting scheme may not accurately reflect the true influence of each active region on full-disk flare prediction probability. It's important to highlight that the weights of these active regions are generally unknown; there are no established methods to accurately determine them, nor are there any prior assumptions that guide the assignment of these weights. Furthermore, the magnetic field measurements, 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 ±60° the magnetic field readings are distorted [5]); therefore, the aggregated full-disk flare probability is in fact, restrictive (i.e., from ARs in central locations) as the data in itself is limited to ARs located within ±45° [6] to ±70° [7] and in some cases, even ±30° [8] due to severe projection effects [9]. This further underscores the inherent challenges in issuing a full-disk flare forecast using an AR-based model.

The full-disk models, however, utilize entire magnetograms corresponding to the full-disk and rely on shape-based features such as size, directionality, sunspot borders (or shapes), and inversion lines similar to the findings from [10]–[13]. These shape-based features in full-disk magnetograms collectively pertain to active regions that are widely recognized as precursors to solar flares. By leveraging convolutional neural networks (CNNs), which are adept at capturing spatial patterns and relationships, the information within magnetograms can be effectively analyzed. The CNN models can automatically extract relevant spatial features and discern the intricate structures and configurations associated with active regions prone to solar flares. This enables the CNNs to recognize and learn the specific spatial characteristics that indicate an elevated probability of solar flare occurrence. In essence, this approach harnesses the power of deep learning for spatial analytics to

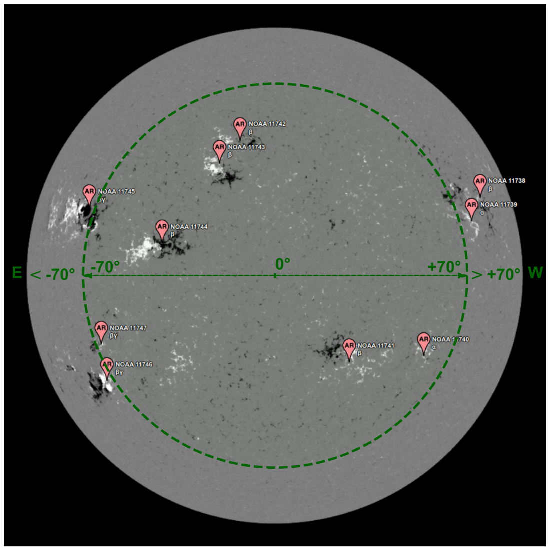


Fig. 1. An annotated full-disk line-of-sight magnetogram observed on 2013-05-13 at 02:00:00 UTC as an example, showing the central location (within ±70°) and near-limb (beyond ±70° to ±90°) region with all the visible active regions present at the noted timestamp, indicated by the red flags. Note that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

interpret the underlying patterns and variations within full-disk line-of-sight (LoS) magnetograms, thereby enabling robust prediction of solar flares. While projection effects persist in these images, it remains to be proven whether full-disk models are capable of predicting flares from areas close to the near-limb. Thus, we provide quantitative evidence favoring a full-disk model and show that it is essential to supplement AR-based models, enabling the prediction of flares in the Sun's near-limb areas and enhancing operational systems.

In recent years, deep learning has emerged as a powerful tool for analyzing and interpreting large volumes of solar data and has shown great experimental success (e.g., [6], [8], [14]–[16]), capturing complex features that precede the onset of solar flares, surpassing traditional statistical methods. However, most of these studies either only use the datasets that correspond to the central locations (within ±70° in longitude of the solar disk, as indicated by an inner-circle region in Fig. 1) and assess their model's efficacy only within these regions, or although they utilize full-disk models, fail to show and discuss the model's prediction capabilities in the near-limb regions (beyond ±70° in longitude of the solar disk). These near-limb regions in full-disk solar magnetograms refer to the areas in the proximity of the edge of the visible disk, which constitute approximately 40% in terms of area of the entire observable solar disk area and is crucial for ultimate operational efforts in the prediction of solar flares. In this work, we explore deep learning for the prediction of ≥M-class solar flares in binary mode with three widely used pre-trained CNNs – AlexNet [17], VGG16 [18], and ResNet34 [19], utilizing hourly sampled instances of full-disk LoS magnetogram images covering solar cycle 24. The focus of this work is to study whether our models can be relied upon for critical applications, particularly in the absence of alternatives, as in the case of near-limb forecasting. Our study shows that the deep learning models can learn the spatial patterns from full-disk magnetogram images, even when the flare originates from the near-limb regions, and provides compelling quantitative evidence supporting the use of a full-disk model as a complement to AR-based models, highlighting its pivotal role in enabling precise prediction of solar flares in near-limb regions. This is a significant addition to existing operational systems and has the potential to greatly enhance space weather forecasting capabilities.

The remainder of the paper is structured as follows. Sec. II provides an overview of existing studies on solar flare predictions using deep learning models and various data sources. In Sec. III, we detail the process of data collection with labeling and consequent data distribution. In Sec. IV we outline our methodology by describing all three architectures explored in this work. Sec. V presents the experimental design and evaluates the effectiveness of our model evaluated with skill scores and prediction sensitivity in central and near-limb regions. Finally, in Sec. VI, we summarize our findings and suggest avenues for future research.

II. Related Work

Solar flare prediction currently relies on four major strategies: (i) empirical human prediction (e.g., [20], [21]), 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) statistical prediction (e.g., [22], [23]), which involves studying the historical behavior of flares to predict their likelihood in the future; (iii) physics-based numerical simulations (e.g., [24], [25]), which involves a detailed understanding of the Sun's magnetic field and the processes that drive flare activity and running simulations models based on the physics of the Sun to predict the occurrence of flares; and (iv) machine learning and deep learning approaches (e.g., [4], [6], [8], [26]–[28]), which involves training algorithms to recognize patterns in solar activity that are associated with flares and using those patterns to make predictions. The rapid progress of machine learning and deep learning techniques has greatly accelerated research efforts in solar flare prediction, offering promising avenues for substantial improvements in forecast accuracy.

The use of machine learning techniques to automatically extract forecast patterns from the intrinsic magnetic field data on the photosphere of the sun has been an active area of research since the early 1990s [29]. Since then, there has been a significant advancement in machine learning and deep learning techniques, leading to a surge of interest in applying these methods to build more accurate flare forecasting models. For instance, in [15], a multi-layer perceptron-based model was employed for predicting ≥C- and ≥M-class flares. The model utilized 79 manually selected physical precursors extracted from multi-modal solar observations, demonstrating the potential of machine learning in flare prediction. Deep learning models have recently emerged as a popular choice for solar flare prediction. In [8], a CNN-based model was trained using solar Active Region (AR) patches extracted from LoS magnetograms within ±30° of the central meridian to predict $\geq$C-, $\geq$M-, and $\geq$X-class flares. Similarly, [6] developed a CNN-based model that issued binary class predictions for $\geq$C- and $\geq$M-class flares within 24 hours using Space-Weather Helioseismic and Magnetic Imager Active Region Patches (SHARP) data [30]. The SHARP data was extracted from solar magnetograms using AR patches located within ±45° of the central meridian. Notably, both models [6], [8] had limited operational capability, as they were restricted to small portions of the observable disk in central locations (±30° and ±45°).

Recently, [27] presented a CNN-based model to predict $\geq$M-class flares using full-disk LoS magnetograms. The model was trained using bi-daily observations (i.e., two magnetograms per day) and achieved a true skill statistic (TSS) of approximately 0.47 and a Heidke skill score (HSS) of approximately 0.35. However, the limited number of instances in the dataset may have affected the model's performance and this study did not investigate the model's ability to predict flares in near-limb regions. Subsequently, [28] developed deep learning-based models that use a similar approach of bi-daily observations of full-disk magnetograms to predict $\geq$C4- and $\geq$M-class flares in binary mode. It is important to note that all the instances that fall between the $\geq$C4- and $\geq$M-class flares were excluded in both training and validation sets. These particular sets of instances lie on the border of two binary outcomes and can be considered the harder-to-predict instances. For the $\geq$M-class flare model, they reported a TSS of $\sim$0.55 and an HSS of $\sim$0.43. However, this study as well did not investigate whether these full-disk models can predict the onset of flares in near-limb regions and no claims were made regarding the model's performance in these regions.

In this work, we build a set of models using compressed full-disk LoS magnetograms with pre-trained deep-learning models to predict the occurrence of $\geq$M-class solar flares. The novel contributions of this paper are as follows: (i) We show an improved overall performance of a full-disk solar flare prediction model by building and comparing three CNN architectures on full-disk magnetogram images, (ii) We provide an extended spatial analysis on the predictive capability of full-disk models on near-limb and central locations, and (iii) We provide results that underscore the pivotal role of full-disk models in the prediction of solar flares in near-limb regions.

III. Data

We used full-disk LoS solar magnetograms obtained from the Helioseismic and Magnetic Imager (HMI) [31] instrument onboard Solar Dynamics Observatory (SDO) [32] available as compressed JPEG 2000 (JP2) images in near real-time publicly via Helioviewer [33]. 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 flare classes) within the next 24 hours, as illustrated in Fig. 2. To elaborate, if the maximum X-ray intensity of a flare was weaker than M, we labeled the observation as "No Flare" (NF: <M), and if it was $\geq$M, we labeled it as "Flare" (FL: $\geq$M). This resulted in 54,649 instances for the NF class and 9,000 instances (8,120 instances of M-class and 880 instances of X-class flares) for the FL class. 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 (solar cycle 24) into four partitions. Partition-1 contained data from January to March, Partition-2 contained data from April to June, Partition-3 contained data from July to September, and Partition-4 contained data from October to December, as shown in Fig. 3. Due to the scarcity of $\geq$M-class flares, the overall distribution of the data is highly imbalanced, with FL:NF $\sim$1:6.

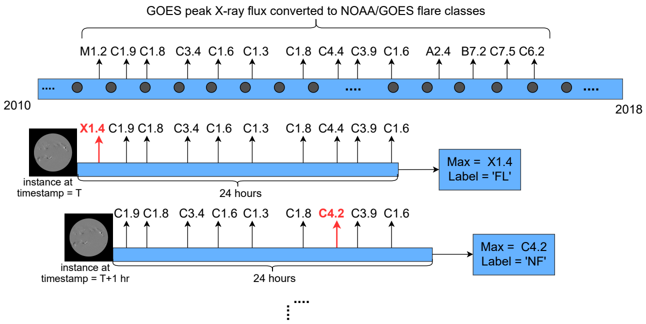

GOES peak X-ray flux converted to NOAA/GOES flare classes
M1.2 C1.9 C1.8 C3.4 C1.6 C1.3 C1.8 C4.4 C3.9 C1.6 A2.4 B7.2 C7.5 C6.2
2010
2018
X1.4 C1.9 C1.8 C3.4 C1.6 C1.3 C1.8 C4.4 C3.9 C1.6
instance at timestamp = T
24 hours
Max = X1.4
Label = 'FL'
C1.9 C1.8 C3.4 C1.6 C1.3 C1.8 C4.2 C3.9 C1.6
instance at timestamp = T+1 hr
24 hours
Max = C4.2
Label = 'NF'

Fig. 2. A visual illustration of the data labeling process using hourly observations of full-disk LoS magnetogram images with a prediction window of 24 hours. Here, 'FL' and 'NF' indicates 'Flare' and 'No Flare' for binary prediction mode ($\geq$M-class flares). The gray-filled circles indicate hourly spaced timestamps for magnetogram instances.

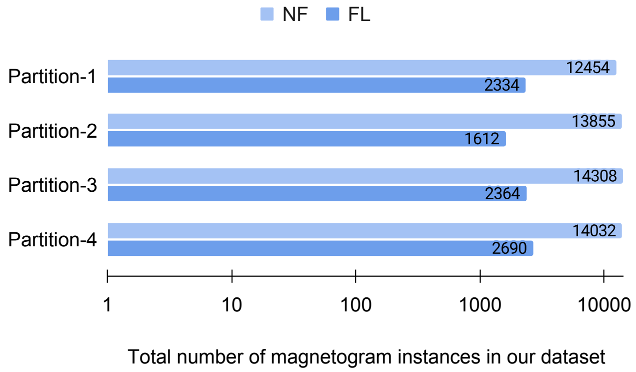


Fig. 3. Data distribution of four tri-monthly partitions for predicting $\geq$M1.0-class flares. Note that the length of the bar is in logarithmic scale.

IV. Models

In this work, we utilize three eminent CNN architectures: AlexNet, VGG16, and ResNet34. Our initial selection was AlexNet [17], a model distinguished by its uncomplicated architecture, which consists of 5 convolutional layers, 3 max pool layers, 1 adaptive average pool layer, and three fully connected layers. The inherent structural simplicity of AlexNet rendered it a desirable candidate for our exploratory analysis. Progressing further, our study included VGG16 [18], a more complex model, to evaluate the hypothesis that an increase in the number of layers might engender enhanced performance. This model augments the foundational structure of AlexNet by integrating additional convolutional layers, all employing uniform 3x3 convolutional kernels. The VGG16 architecture encompasses 13 convolutional layers, 5 max pool layers, 1 adaptive average pool layer, and 3 fully connected layers. Lastly, we included ResNet34 [19], a CNN model that extends the complexity of the VGG16 design by facilitating the training of deeper networks with fewer parameters. Diverging from the methodologies employed by AlexNet and VGG16, ResNet34 integrates residual connections from each layer into subsequent connected layers. The architecture of ResNet34 consists of 33 convolutional layers, including a 7x7 kernel for the initial layer and 3x3 kernels for the remaining layers, along with one max pool layer, one adaptive average pool layer, and one fully connected layer. The primary reason behind our choice of these distinct architectures was to analyze and evaluate the influence of varying architectural designs and increasing layer depths on performance. Additionally, we factored the simplicity of the architectures into our selection process, in light of the relatively modest scale of our dataset suitable for deep learning models. These pre-trained models require a 3-channel image for input, however, our data comprises compressed solar magnetogram images, which are grayscale. To reconcile this, we incorporated an additional convolutional layer at the onset of the network architecture as shown in Fig. 4, which accepts a 1-channel input. This layer employs a 3×3 kernel with a size-1 stride, padding, and dilation, and consequently generates a 3-channel image. This added convolutional layer is initialized using Kaiming initialization [34] for all three models. Additionally, with the aim of optimally utilizing the pre-trained weights—–irrespective of the architectural specifics of these models, which anticipate 3-channel input of varying dimensions—–-we used an adaptive average pooling layer within each model. This layer is positioned after the completion of feature extraction via the convolutional layer and immediately preceding the fully-connected layer. This placement facilitates the alignment of dimensions with our image input size, which is 512×512.

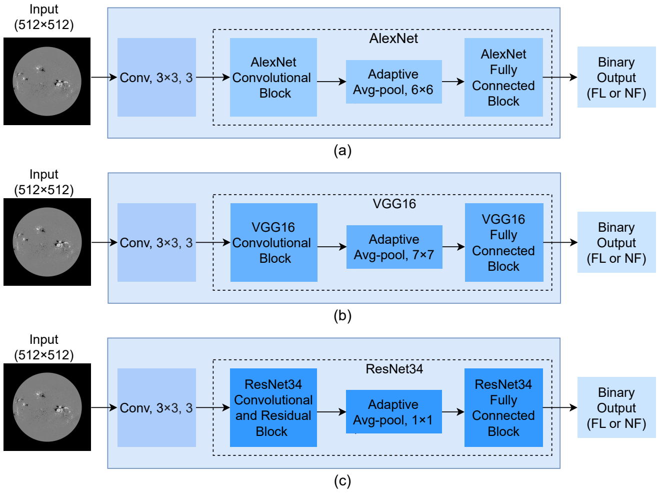


Fig. 4. An overview of three deep learning architectures we use (a) AlexNet-, (b) VGG16-, (c) ResNet34-based models.

V. Experimental Evaluation

In this section, we provide a comprehensive overview of our experimental setup, outlining the settings for data augmentation techniques employed for data balancing, and the hyperparameter configurations utilized to train our models. Moreover, we present the obtained results and share our observational remarks derived from the experiments, with a specific emphasis on the crucial aspect of flare spatial locations, specifically near-limb flares. These near-limb flares are often overlooked, and our analysis sheds light on the predictive capabilities of our models in operational systems.

A. Experimental Settings

We trained our full-disk flare prediction models using Stochastic Gradient Descent (SGD) as the optimizer and Negative Log-Likelihood (NLL) as the objective function. We initialized each of the models with their corresponding pre-trained weights, then further trained it for 50 epochs while employing a dynamic learning rate scheduling strategy, OneCycleLR [35], details presented in Table I. All three models in this study were trained using the OneCycleLR scheduler with cosine annealing, which offers the benefit of automating the learning rate schedule selection for hyperparameter tuning. This scheduler adjusts the learning rate in a cyclical pattern, gradually increasing it to help the model quickly converge and then decreasing it to fine-tune performance. The steps per epoch were set to the number of batches in training data, and the batch size was 64. Utilizing the OneCycleLR scheduler, the models benefit from an automated and optimized learning rate schedule, simplifying the process of hyperparameter tuning.

Building upon the discussion in Sec. III, it is important to acknowledge that our dataset has an inherent class imbalance issue. This imbalance can significantly influence the performance of the models, potentially leading to less precise and reliable predictions for the minority class. To address this, we employed data augmentation and adjusted class weights in the loss function. Specifically, we applied three augmentation techniques: vertical flipping, horizontal flipping, and rotations between +5° and -5° to both classes. For each instance in the minority class (FL), we applied all three augmentations, quadrupling the total number of instances for the entire FL-class. For each instance in NF-class, we randomly selected one of the three aforementioned augmentation techniques, doubling the total instances for this class. The goal of augmenting the NF-class instances was to ensure that the NF-class, though not uniformly augmented, retained a diversity in its data akin to the FL-class and expand the overall dataset. Post augmentation, we adjusted the class weights to be inversely proportional to

TABLE I
PARAMETER SETTINGS FOR ALL OF OUR MODELS: ALEXNET, VGG16, AND RESNET34.

Models	Optimizer	Loss Function	Initial Learning Rate	Max. Learning Rate	Batch Size	Weight Decay	Epochs
AlexNet	SGD	NLL	$1e-5$	$1e-4$	64	$1e-4$	50
VGG16	SGD	NLL	$1e-5$	$1e-5$	64	$1e-4$	50
ResNet34	SGD	NLL	$1e-5$	$1e-5$	64	$1e-3$	50

class frequencies, thereby penalizing misclassifications of the minority class. Finally, we evaluated our models using a 4-fold cross-validation approach on tri-monthly partitions.

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

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

TSS and HSS values range from -1 to 1, where 1 indicates all correct predictions, -1 represents all incorrect predictions, and 0 represents no skill. In contrast to TSS, HSS is an imbalance-aware metric, and it is common practice to use HSS in combination with TSS for the solar flare prediction models due to the high class-imbalance ratio present in the datasets. For a balanced dataset, these metrics are equivalent [36]. In solar flare prediction, TSS and HSS are the preferred choice of evaluation metrics compared to commonly used metrics in image classification, such as accuracy, precision, recall, and F1 scores as they ensure a comprehensive and reliable evaluation of predictive capabilities, especially in scenarios with imbalanced class distributions. Lastly, we report the subclass and overall recall (shown in Eq. 3) for flaring instances (M- and X-class) to assess the prediction sensitivity of our models in central and near-limb regions.

B. Evaluation

This section presents an analysis of the results, focusing on the performance comparison of the models. The findings reveal that the AlexNet-based model exhibits better performance in relation to both the VGG16- and ResNet34-based models, as evidenced by the HSS and TSS scores provided in Table II. Notably, the AlexNet-based model demonstrates enhanced robustness, as indicated by the lower standard deviation, and achieves an approximate 2% improvement (for both TSS and HSS) compared to the VGG16-based model. Furthermore, when compared to the ResNet34-based model, the AlexNet-based model showcases a 1% higher skill score (for both TSS and HSS). It is important to highlight that the skill scores of the VGG16 and ResNet34 models exhibit greater variability, primarily influenced by the outcomes of Fold-3 in the 4-fold cross-validation experiment, details presented in Table. IV. Furthermore, our best results surpass the performance reported in [27] by $\sim 5\%$ in terms of TSS (reported as 0.47±0.06) and by $\sim$2% in terms of HSS (reported as 0.35±0.05)[1].

TABLE II
THIS TABLE PRESENTS THE AGGREGATED RESULTS OF A 4-FOLD CROSS-VALIDATION EXPERIMENT ON OUR MODELS. THE MODELS' AVERAGE PERFORMANCE IS EVALUATED USING TWO SKILL SCORES (TSS AND HSS), BASED ON THE TEST SET.

Models	TSS	HSS
AlexNet	**0.526±0.05**	**0.372±0.05**
VGG16	0.506±0.09	0.353±0.09
ResNet34	0.513±0.09	0.360±0.09

In addition, we evaluate the results by examining the correct prediction and missed flare counts for class-specific flares (X-class and M-class) in central locations and near-limb locations of the Sun, as presented in Table III. It is noteworthy that, while the overall performance measured in terms of TSS and HSS indicates the better performance of the AlexNet-based model over the other two deeper and more advanced models, VGG16 and ResNet34, the ResNet34-based model exhibits the best performance on average for near-limb events. The class-specific analysis for X- and M-class flares reveals that the ResNet34-based model achieves correct predictions for $\sim$81% of the X-class flares ($\sim$16% higher than AlexNet) and $\sim$56% of the M-class flares ($\sim$1% higher than AlexNet) in near-limb locations. Despite all models being fine-tuned with the same dataset and undergoing similar hyperparameter optimization, their distinctive architectures influenced their ability to capture specific spatial patterns and features, leading to variations in overall performance, prediction sensitivity, and recall rates for specific flare intensities and event locations.

[1] While there are several other work (mentioned in Sec. II) in solar flare prediction that evaluate the performance of their deep learning models using TSS and HSS, these models are not directly comparable since they employ different datasets, data timelines, and data partitioning strategies.

TABLE III

A COMPREHENSIVE OVERVIEW OF CORRECTLY (TP) AND INCORRECTLY (FN) CLASSIFIED X- AND M-CLASS FLARES, DISTINGUISHING BETWEEN CENTRAL ($|longitude| \leq \pm 70°$) AND NEAR-LIMB LOCATIONS. RECALL STATISTICS ACROSS THESE LOCATION GROUPS ARE ALSO PROVIDED. ALL COUNTS ARE AGGREGATED ACROSS THE FOLDS OF THE CROSS-VALIDATION EXPERIMENT.

		Within ±70°			Beyond ±70°		
Models	Flare-Class	TP	FN	Recall	TP	FN	Recall
AlexNet	X-Class	614	54	**0.92**	138	74	0.65
	M-Class	4,645	1,185	**0.80**	1,276	1,014	0.55
	Total (X&M)	5,259	1,239	**0.81**	1,414	1,088	0.57
VGG16	X-Class	560	108	0.84	165	47	0.78
	M-Class	4,473	1,357	0.77	1,273	1,017	0.55
	Total (X&M)	5,033	1,465	0.77	1,438	1,064	0.57
ResNet34	X-Class	612	56	**0.92**	172	40	**0.81**
	M-Class	4,449	1,381	0.76	1,291	999	**0.56**
	Total (X&M)	5,061	1,437	0.78	1,463	1,039	**0.59**

TABLE IV

DETAILED RESULTS OF 4-FOLD CROSS-VALIDATION EXPERIMENTS, SHOWING ALL THE FOUR OUTCOMES OF CONFUSION MATRICES (TP, FP, TN, FN) FOR ALL THREE MODELS USED IN THIS STUDY.

Models	Folds	TP	FP	TN	FN	TSS	HSS
AlexNet	Fold-1	1,729	2,225	10,229	605	0.5621	0.4385
	Fold-2	1,075	2,298	11,557	537	0.5010	0.3380
	Fold-3	1,660	3,291	11,017	704	0.4722	0.3241
	Fold-4	2,209	3,549	10,483	481	0.5683	0.3890
VGG16	Fold-1	1,704	2,067	10,387	630	0.5641	0.4512
	Fold-2	1,233	3,401	10,454	379	0.5194	0.2841
	Fold-3	1,409	3,089	11,219	955	0.3801	0.2761
	Fold-4	2,125	3,236	10,796	565	0.5593	0.3992
ResNet34	Fold-1	1,779	2,145	10,309	555	0.5900	0.4621
	Fold-2	1,257	3,872	9,983	355	0.5003	0.2550
	Fold-3	1,328	2,299	12,009	1,036	0.4011	0.3280
	Fold-4	2,160	3,382	10,650	530	0.5620	0.3933

Moreover, we scrutinized the proficiency of our models both from a quantitative and qualitative standpoint by conducting an intricate spatial analysis of their performance in correlation with the locations of M- and X-class solar flares, that were used as the labels. For the purpose of our analysis, we used the predictions made on the test set and created a heatmap by gathering the flares grouped by their location in the Heliographic Stonyhurst (HGS) coordinate system, where each bin represents a 5° × 5° spatial cell in terms of latitude and longitude. Initially, we computed the recall for the ≥M-class flares (combined M- and X-class flares) in each spatial cell, providing a comprehensive assessment of the models' performance. Subsequently, we evaluated the recall separately for M-class and X-class flares, allowing us to analyze the models' sensitivity at a more granular level. The heatmaps that illustrate the spatial distribution of recall scores for ≥M-, X-, and M-class flares are shown in Fig. 5, 6 (a) , and 6 (b) respectively. This allowed us to compare all three models on their capabilities to learn spatial patterns that pinpoint the locations where the models were more effective in making accurate predictions and vice versa.

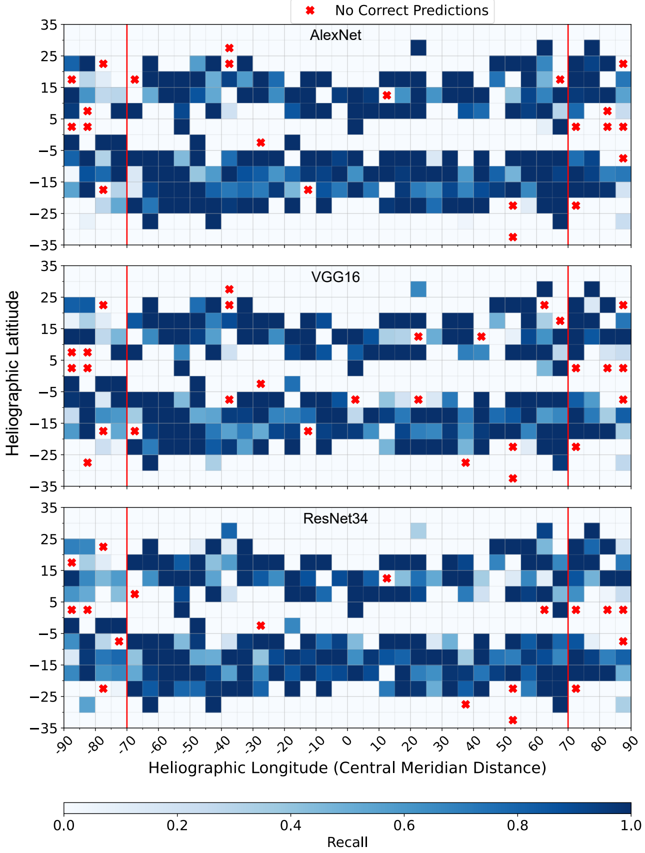
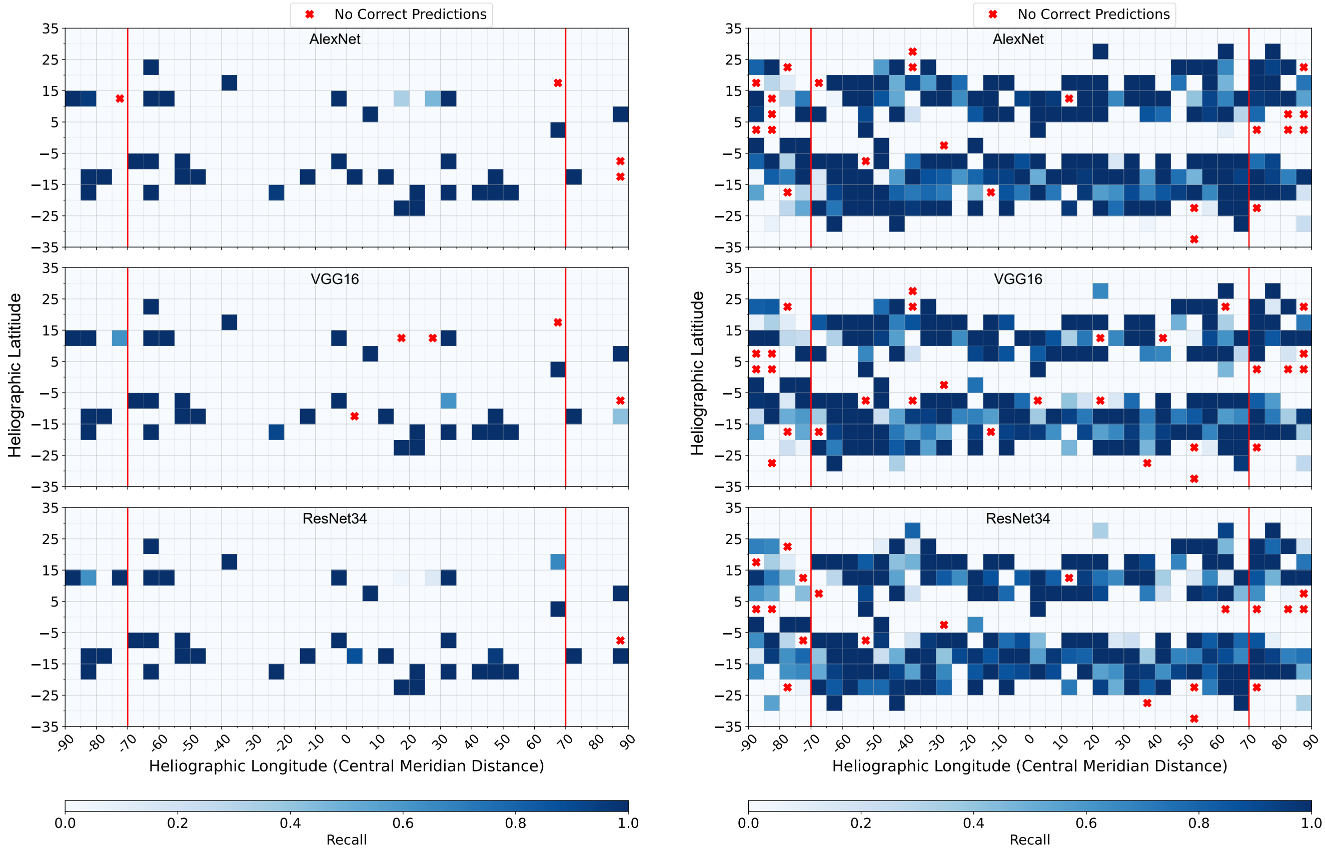


Fig. 5. A heatmap illustrating the quantitative and qualitative evaluation of all three models' recall performance for **≥M-class flares** i.e., FL-class. The locations of the flares (with maximum peak x-ray flux, used as labels) are aggregated into 5° × 5° spatial bins of latitude and longitude. Note: Red cross in white grids represents locations with zero correct predictions while white cells without red cross represent unavailable instances.

Our findings indicate that all three models demonstrated reasonable proficiency in predicting X-class flares in central locations. However, among these, the ResNet34-based model stood out for its overall better performance in accurately forecasting X-class flares, regardless of whether they were in near-limb or central locations as shown in Fig. 6 (a). Upon analysis of the heatmaps for ≥M- and only M-class flares, as depicted in Fig. 5 and Fig. 6 (b) respectively, it was revealed that the ResNet34-based model generally yielded more accurate predictions across diverse spatial locations in comparison to the other models. Nonetheless, a common limitation across all three models was an elevated rate of false negatives in near-limb areas for M-class flares. Notably, these regions are often associated with unreliable readings due to projection effects. Despite this challenge, our study



(a)

(b)

Fig. 6. A heatmap illustrating the quantitative and qualitative evaluation of all three models' recall performance for (a) **X-class flares** and (b) **M-class flares**. The locations of the flares (with maximum peak x-ray flux, used as labels) are aggregated into 5° × 5° spatial bins of latitude and longitude. Note: Red cross in white grids represents locations with zero correct predictions while white cells without red cross represent unavailable instances.

signifies a substantial progression in space weather forecasting, enabling the prediction of flares even in these intricate near-limb regions with distorted magnetic fields. The ability to accurately identify flare locations could considerably enhance the precision of operational forecasting methods. This ability to predict flares in traditionally overlooked near-limb areas has considerable implications.

VI. Conclusion and Future Work

In conclusion, our study involved the development and evaluation of deep learning models, namely AlexNet, VGG16, and ResNet34, for the prediction of solar flares, with a specific emphasis on capturing near-limb events. Through rigorous analysis and examination of the results, several significant findings have emerged, providing insights into the capabilities of these models. Firstly, our models demonstrated promising abilities in learning intricate spatial patterns from full-disk magnetograms, illustrating the potential of deep learning techniques in extracting meaningful features from complex solar data. Of particular importance, our investigation unveiled a notable performance advantage of the ResNet34-based model in predicting near-limb flares. This finding highlights the efficacy of employing deeper architectures with residual connections, which enhance feature extraction and facilitate the capture of subtle patterns associated with near-limb events. Moreover, our study highlighted the variability in model performance across different flare types and event locations, emphasizing the importance of tailoring models and analyzing results in context-specific manners. This underlines the need for further exploration of model architecture enhancements and training techniques to effectively capture the diverse nature of flare events. The implications of our research extend to operational forecasting systems, where the precise and reliable prediction of solar flares, including near-limb events, holds significant importance. The improved capabilities demonstrated by our models provide valuable insights for refining forecasting methodologies in operation and facilitating real-time decision-making processes.

Apart from the promising capabilities of our models, it is important to highlight the associated inherent challenges. Factors such as data availability, observational constraints, and the evolving nature of solar activity pose ongoing obstacles to model development and validation. Addressing these challenges necessitates advancements in data collection, integration, and the developing of sophisticated models. Future research directions can explore the integration of multi-modal data, the development of models that can capture temporally evolving solar activity, the interpretability of learned features, and the utilization of explainable deep learning techniques to enhance predictive capabilities and address limitations. Overall, our study contributes to the growing body of research in solar flare prediction, shedding light on the capabilities and limitations of different model architectures, particularly for near-limb flares. These insights hold the potential to drive advancements in ultimate operational forecasting efforts.

References

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

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

[3] K. D. Leka, S.-H. Park, K. Kusano, J. Andries, G. Barnes, S. Bingham, D. S. Bloomfield, A. E. McCloskey, V. Delouille, D. Falconer, P. T. Gallagher, M. K. Georgoulis, Y. Kubo, K. Lee, S. Lee, V. Lobzin, J. Mun, S. A. Murray, T. A. M. H. Nageem, R. Qahwaji, M. Sharpe, R. A. Steenburgh, G. Steward, and M. Terkildsen, "A comparison of flare forecasting methods. II. benchmarks, metrics, and performance results for operational solar flare forecasting systems," *The Astrophysical Journal Supplement Series*, vol. 243, no. 2, p. 36, Aug. 2019.

[4] 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.

[5] D. A. Falconer, S. K. Tiwari, R. L. Moore, and I. Khazanov, "A new method to quantify and reduce the net projection error in whole-solar-active-region parameters measured from vector magnetograms," *The Astrophysical Journal*, vol. 833, no. 2, p. L31, Dec. 2016.

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

[7] A. Ji, B. Aydin, M. K. Georgoulis, and R. Angryk, "All-clear flare prediction using interval-based time series classifiers," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2020, pp. 4218–4225.

[8] X. Huang, H. Wang, L. Xu, J. Liu, R. Li, and X. Dai, "Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms," *The Astrophysical Journal*, vol. 856, no. 1, p. 7, Mar. 2018.

[9] 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.

[10] M. B. Korsós, T. Baranyi, and A. Ludmány, "PRE-FLARE DYNAMICS OF SUNSPOT GROUPS," *The Astrophysical Journal*, vol. 789, no. 2, p. 107, Jun. 2014.

[11] A. E. McCloskey, P. T. Gallagher, and D. S. Bloomfield, "Flaring rates and the evolution of sunspot group McIntosh classifications," *Solar Physics*, vol. 291, no. 6, pp. 1711–1738, Jun. 2016.

[12] V. Deshmukh, T. E. Berger, E. Bradley, and J. D. Meiss, "Leveraging the mathematics of shape for solar magnetic eruption prediction," *Journal of Space Weather and Space Climate*, vol. 10, p. 13, 2020.

[13] A. Ji, X. Cai, N. Khasayeva, M. K. Georgoulis, P. C. Martens, R. A. Angryk, and B. Aydin, "A systematic magnetic polarity inversion line data set from SDO/HMI magnetograms," *The Astrophysical Journal Supplement Series*, vol. 265, no. 1, p. 28, Mar. 2023.

[14] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, S. Watari, and M. Ishii, "Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms," *The Astrophysical Journal*, vol. 835, no. 2, p. 156, jan 2017.

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

[16] 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.

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

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

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

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] M. B. Korsós, M. K. Georgoulis, N. Gyenge, S. K. Bisoi, S. Yu, S. Poedts, C. J. Nelson, J. Liu, Y. Yan, and R. Erdélyi, "Solar flare prediction using magnetic field diagnostics above the photosphere," *The Astrophysical Journal*, vol. 896, no. 2, p. 119, Jun. 2020.

[26] M. G. Bobra and S. Couvidat, "Solar flare prediction using≤SDO≤/HMI VECTOR MAGNETIC FIELD DATA WITH a MACHINE-LEARNING ALGORITHM," *The Astrophysical Journal*, vol. 798, no. 2, p. 135, Jan. 2015.

[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] T. Aso, T. Ogawa, and M. Abe, "Application of back-propagation neural computing for the short-term prediction of solar flares." *Journal of geomagnetism and geoelectricity*, vol. 46, no. 8, pp. 663–668, 1994.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," 2017.

[36] 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 Astrophysical Journal Supplement Series*, vol. 254, no. 2, p. 23, May 2021.