Deep Neural Networks Based Solar Flare Prediction Using Compressed Full-disk Line-of-sight Magnetograms

Chetraj Pandey(✉), Rafal A. Angryk, and Berkay Aydin

Georgia State University, Atlanta, GA 30302, USA
{cpandey1,rangryk,baydin2}@gsu.edu

Abstract. The efforts in solar flare prediction have been engendered by the advancements in machine learning and deep learning methods. We present a new approach to flare prediction using full-disk compressed magnetogram images with Convolutional Neural Networks. We selected three prediction modes, among which two are binary for predicting the occurrence of ≥M1.0 and ≥C4.0 class flares and one is a multi-class mode for predicting the occurrence of <C4.0, [≥C4.0, <M1.0] and ≥M1.0 within the next 24 h. We perform our experiments in all three modes using three well-known pretrained CNN models—AlexNet, VGG16 and ResNet34. For this, we collect compressed 8-bit images derived from full-disk line-of-sight magnetograms provided by the Helioseismic and Magnetic Imager (HMI) instrument onboard Solar Dynamics Observatory (SDO). We trained our models using data-augmented oversampling to address the existing class-imbalance issue by following a time-segmented cross-validation strategy to effectively understand the accuracy performance of our models and used true skill statistics (TSS) and Heidke skill score (HSS) as metrics to compare and evaluate. The major results of this study are (1) we successfully implemented an efficient and effective full-disk flare predictor for operational forecasting using compressed images of solar magnetograms; (2) Our candidate model for multi-class flare prediction achieves an average TSS of 0.36 and average HSS of 0.31. Similarly, for binary prediction in (i) ≥C4.0 mode: we achieve an average TSS score of 0.47 and HSS score of 0.46, (ii) ≥M1.0 mode: we achieve an average TSS score of 0.55 and HSS score of 0.43.

Keywords: Solar flares · Deep neural networks · Solar magnetograms

1 Introduction

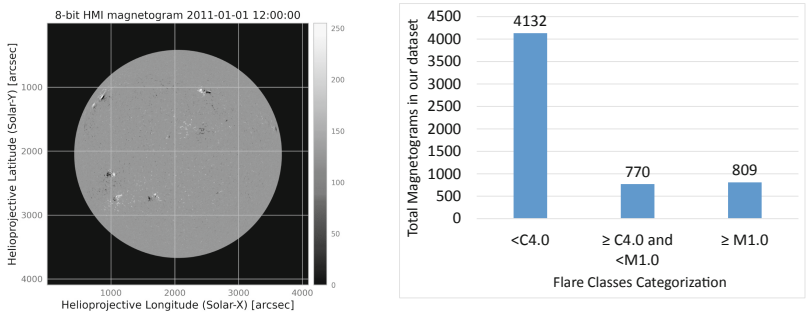
Solar flares are the large eruptions of electromagnetic radiation originating from the inner solar atmosphere and extending out to the outermost atmosphere of the Sun, which can last minutes to hours, and they often transpire as a sudden flash of increased brightness on the Sun observed near its surface [1]. Although there exists observational precursors, the actual physical cause of this phenomenon is

J. A. Lossio-Ventura et al. (Eds.): SIMBig 2021, CCIS 1577, pp. 380–396, 2022.
https://doi.org/10.1007/978-3-031-04447-2_26

still unsolved, which hinders the validation process of statistical or data-driven flare forecasts. Recent studies [2,3] shows promising results when solar flare prediction is posed as a computer vision/image classification task and deep architectures are employed.

Solar flares are categorized into five major classes according to their peak X-ray flux level : X ($> 10^{-4}\mathrm{Wm}^{-2}$), M ($> 10^{-5}\mathrm{Wm}^{-2}$), C ($> 10^{-6}\mathrm{Wm}^{-2}$), B ($> 10^{-7}\mathrm{Wm}^{-2}$), and A ($> 10^{-8}\mathrm{Wm}^{-2}$) [4]. These flare classes are measured in logarithmic scales (i.e., M3.2 is 10 times stronger than C3.2 flare). Although, the explosive heat of a solar flare cannot reach all the way to the Earth, the electromagnetic radiation and energetic particles certainly can induce the intense variation in near-Earth magnetic field, causing potential disruptions to many stakeholders such as the electricity supply chain, airlines industry, astronauts in space and communication systems including satellites and radio. The X-class and M-class flares are rare events and hence the scarcity of data give rise to the class-imbalance issue which further complicates the learning process for deep learning models, where the large amount of data is considered to be crucial for achieving meaningful generalization.

Most of the current flare prediction models are active region-based, that is, predictions are issued for a certain region on the Sun. Active regions are the temporary areas on the Sun characterized by especially strong and complex magnetic fields. These regions frequently produce various types of solar activity and are well-suited for predicting the occurrence of flares. For an operational system—system which is ready to use with the near real-time data for making real-time predictions—individual predictions from active regions are aggregated to provide a final prediction result. However, due to the strong projection effects near the limbs of the Sun, such predictions are limited to the active regions in



(a) (b)

Fig. 1. (a) A pictorial representation of the compressed 8-bit image derived from the line-of-sight magnetogram as observed by SDO/HMI on 2011–01-01 12:00:00 UT. (b) The total number magnetograms for each target class label we use in this study; i.e., <C4.0 class, ≥C4.0 to <M1.0 and ≥M1.0 class flares.

central locations, which is not ideal for operational systems. Full-disk predictions are therefore more appropriate to complement the active region-based counterparts and provide a crucial, often overlooked, element to these near real-time operational systems.

Convolutional Neural Networks (CNN) [5] based deep learning architectures have been very popular for over a decade now for computer vision problems where data are labeled images. In this experiment, we use 8-bit compressed magnetogram images where the pixel value ranges from 0 to 255 derived from full-disk line-of sight solar magnetograms which contains 4096 × 4096 raster map of the one dimensional magnetic field strength values on the sun typically ranging from $\sim \pm 4500$G. Using compressed images instead of high depth solar magnetograms do not show any reliable magnetic field information however, they represent the shape parameters of active-regions which includes the projected shape of sunspot at an angle. Considering the limited scope of active region-based flare prediction counterparts, where the prediction is limited to central location (up to $\pm 70°$) of a full-disk magnetogram due to severe projection effect on the limbs of the Sun, with compressed full-disk magnetograms we incorporate the entire information including the active-regions present on the limbs. Although the 8-bit compressed magnetograms may induce information loss to some extent; however, considering the depth and complexity of deep learning models, it may be a more suitable choice to use images as it elevates the model's computational efficiency while training and predicting the flaring events in real-time.

In this paper, we address the task of training robust full-disk flare prediction models and explore different prediction modes (i.e., predicting the occurrence of $\geq$C4.0 and $\geq$M1.0 class flares in binary mode, and $<$C4.0, $\geq$C4.0 to $<$M1.0 and $\geq$M1.0 in multi-class mode with a prediction window of 24 h) and assess the impact of such formulation of the prediction problem with three different CNN architectures. As mentioned earlier, we use compressed images of full-disk line-of-sight magnetograms obtained from Helioseismic and Magnetic Imager (HMI) onboard Solar Dynamics Observatory (SDO). These images do not require further preprocessing and are available in near real-time (often $<$30 min). An example compressed solar magnetogram image is demonstrated in Fig. 1.(a). These compressed images contain 4096 × 4096 pixels which we resize to 512×512 pixels for our experiments. We use a transfer learning based approach with three landmark CNN models, AlexNet [6], VGG16 [7] and ResNet34 [8]. We customize these models as per our requirement of two classes in binary mode and three classes in multi-class mode with single-channel input image and analyze the performance of each model.

In the long run, we intend to employ these models to create more reliable and robust flare prediction ensembles in an operational setting. Robust prediction of solar flares is a central problem in space weather forecasting and has many practical implications. Many of the severe solar storms are associated with a strong-flare and deep learning-based prediction models have the potential to help understand intrinsic magnetic field configurations that lead to a flare. We also note that the models trained in our work are not active region-based and they only use data derived from line-of-sight magnetograms.

The remainder of this paper is organized as follows. In Sect. 2, we present the related work on solar flare predictions using machine and deep learning models. In Sect. 3, the data collection and preparation strategies are presented. In Sect. 4, we present the overview of the model architectures we use for solar flare prediction. In Sect. 5, we present our detailed experimental evaluation, and, lastly, in Sect. 6, we present our final remarks on this work including its limitations and discuss future work.

2 Related Work

The convolutional neural network (CNN) [5] is a class of deep neural network architecture with sparse neuron connections inspired by biological processes [9] to imitate the animal visual cortex. Recently, there have been several attempts to predict solar flares using deep learning models. Nishizuka et al. presented a Deep learning model based on a multi-layer perceptron for solar flare forecasts for ≥C1.0 and ≥M1.0 class [10]. In this study, they used 79 manually selected features (well-known physical precursors) extracted from multi-modal solar observations, which are vector magnetograms, 131 Å AIA images, and 1600 Å UV continuum images. Their models require a preliminary feature extraction process to prepare the data to feed the deep learning model.

Similarly, Huang et al. [11] presented a CNN-based flare forecasting model with two convolutional layers with 64 11×11 kernels where they used solar active regions patches extracted from line-of-sight solar magnetograms within ±30° of the central meridian. In this work, their models are trained to predict ≥C1.0-, ≥M1.0-, and ≥X1.0-class flares from active regions in central locations. While they show significantly high accuracy (>0.66 true skill statistic) for ≥M1.0 class, the models are limited only to certain areas of the observable disk, overlooking the significant portion that has information on other active-regions, and thus have limited operational prediction ability. In [2], Park et al. applied a CNN-based hybrid model which combines GoogleLeNet [12] and DenseNet [13]. Their model is trained to predict the occurrence of a ≥C1.0 class within the next 24 h. They use data from both HMI magnetograms, as well as magnetograms from Michelson Doppler Imager (MDI) onboard Solar and Heliospheric Observatory (SOHO), the predecessor of HMI/SDO. This allowed them to use a substantially higher number of images for training (entire MDI dataset, one image per day, for training and HMI dataset for testing); however, it should be noted that these two instruments are currently not cross-calibrated for use in forecasting and may lead to spurious or deficient patterns being discovered.

Li et al., in [3], also use a CNN-based model to issue binary class predictions for both ≥C1.0 class and ≥M1.0 class flares within 24 h using Space-Weather Helioseismic and Magnetic Imager Active Region Patches (SHARP) data [14] extracted from solar magnetograms located within ±45° of the central meridian excluding the magnetograms samples that has multiple sunspot groupings (or NOAA-defined active regions). This again limits the scope of the prediction to easier-to-predict active regions. They use undersampling and data augmentation

to remedy the class-imbalance issue and create a non-chronological dataset by randomizing the process of data splitting for a 10-fold cross validation. While such data splitting leads to higher experimental accuracy scores, it often fails to deliver similar real-time performance as discussed in [15].

In this work, we build a set of models using compressed full-disk line-of-sight magnetograms with pretrained deep learning models to predict the occurrence of flaring events (for ≥C4.0 and ≥M1.0 class in binary modes and <C4.0, ≥C4.0 to <M1.0, and ≥M1.0 class in multi-class mode) with a prediction window of 24 h. We will use bi-daily full-disk images sampled at 00:00 UT and 12:00 UT, and labeled based on the existence of a flaring event within the next 24 h. For this, we create a dataset by using a non-chronological splitting of data into four time segmented partitions for both binary and multi-class flare predictions. We use 8-bit compressed images of full-disk line-of-sight solar magnetograms with a modified version of the pretrained AlexNet, VGG16 and ResNet34 models for all of our experiments. To remedy the existing class-imbalance issue in the dataset we use data-augmented oversampling.

3 Data Preparation

We use an image dataset derived from full-disk line-of-sight HMI solar magnetograms. HMI provides various magnetic field products at high spatial and temporal resolution. We select two images derived from magnetograms at 00:00 UT and 12:00 UT each day from December 2010 to December 2018. These images are not the original full-depth magnetic field rasters but rather are compressed JP2 images created from magnetograms (i.e., pixel values ranging from 0–255). We retrieve our images from a public data API, Helioviewer [16], which provides 4096 × 4096 compressed images of magnetograms closest to the requested timestamp. While preparing our final dataset, we skip the timestamp if the observation time of the available image and requested image timestamp is more than six hours.

We use a prediction window (i.e., forecast horizon) of 24 h. The bi-daily observations of magnetograms are labeled based on the maximum of peak X-ray flux within the next 24 h, converted to GOES flare classes; e.g., if the maximum intensity flare for the next 24 h (starting from the image's observation time) is an M1.2 flare, then we tentatively label the image as 'M'.

We collect a total of 5,711 solar magnetograms where there are 81 X-class flares, 728 M-class flares, 2,324 C-class flares, and 2,578 are <C1.0[1]. To perform the task of multi-class flare prediction we choose a threshold of C4.0 where flare <C4.0 are considered to be flare-quiet instances and ≥C4.0 class are further subdivided into two flaring classes. The main motivation to choose this threshold is that in most cases, the flares above C4.0 are observed to be associated

[1] While there may be A-, B- and lower C-class flares in our $< C4.0$ category, they are often referred to as *flare-quiet* (or *no-flare*) category, because these flares are weak and may not be detected properly during solar maxima due to high background X-ray flux.

with Coronal Mass Ejection (CME) events with notably higher speed that can make impact on the near-Earth space. Furthermore, using C4.0 as the threshold ensures approximately equal number of instances i.e. 770 and 809 images for two flaring classes ($\geq$C4.0 to $<$M1.0 and $\geq$M1.0) and we refer to them as mild-flares and strong-flares respectively in the scope of this paper as shown in Fig. 1.(b).

When preparing the dataset for $\geq$C4.0 class in binary prediction mode, if the maximum X-ray intensity of flare is weaker than C4.0 ($<$C4.0), the observations are labeled as "no-flare" and greater than or equal to ($\geq$C4.0) are labeled as "flare". In doing so, we collect 4,132 "no-flare" instances and 1579 "flare" instances. Similarly for $\geq$M1.0 class flares prediction in binary prediction mode, we do not include mild-flares ($\geq$C4.0 to $<$M1.0) to train our models. The objective for excluding those instances is to make the decision boundary for $\geq$M1.0 class wider so that the model could generalize better. For this, we collect 4,132 "no-Flare" instances and 809 "flare" instances for $\geq$M1.0 class binary prediction mode.

As we will describe later in the experimental evaluation, we create our cross-validation (CV) dataset partitions based on the tri-monthly partitioning of total images. The average class-imbalance ratio in our entire dataset for binary prediction in $\geq$C4.0 class mode is $\sim$1:2.6 (flare:no-flare). On the other hand, due to scarcity of X- and M- class flares, for $\geq$M1.0 class flares, after excluding the mild-flares from no-flare instances, the data distribution is highly imbalanced, $\sim$1:5 (flare:no-flare). Similarly for multi-class prediction, the two of the flaring classes (mild and strong) are nearly balanced, but no-flare class is still the majority class. The imbalance ratio is $\sim$1:1:5 (strong-flare:mild-flare:no-flare).

4 Model Architecture

A general architecture of a CNN model in a classification problem consists of convolutional layers with *ReLU* activation function followed by a pooling layer and finally one or more fully connected layers with a *softmax* function to give the prediction probabilities of each class [17]. In CNNs, each convolutional layer has a set of kernels (filters), which are trained to extract complex features from the input data. After the convolutional layer, we use ReLU activation which adds non-linearity to the model. To summarize the outputs from a convolutional layer by reducing the size of the output map, a pooling layer is used. Pooling layers maximize or average the spatial size of output from the convolutional layer and reduce the number of computations. A fully connected layer is the traditional neural network where nodes in one layer are densely connected with nodes in another fully connected layer. To overcome the problem of overfitting in such deep networks, usually a dropout layer [18] is added, which ensures the random sparse connectivity between the nodes in two fully connected layers [19].

In this study, we implement three of the well-known CNN-architectures: AlexNet, VGG16 and ResNet34 models to make binary and multi-class flare predictions. In the first place, we use AlexNet model [6] because of it's simplicity in the architecture which consists of 5 convolutional layers, 3 maxpool layers,

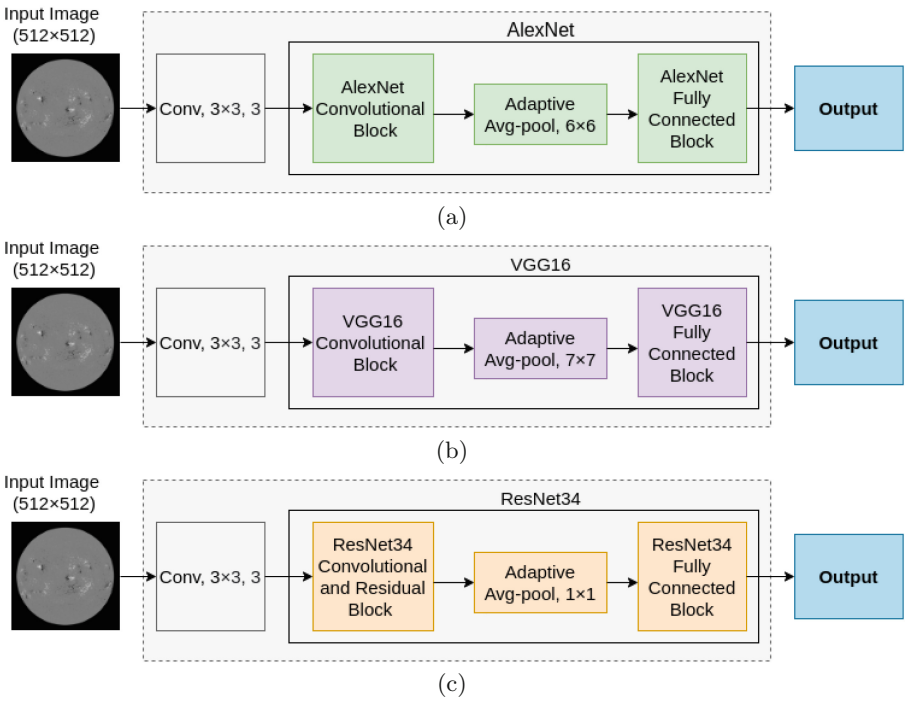


Fig. 2. An overview of three deep learning architectures we use (a) AlexNet-, (b) VGG16-, (c) ResNet34-based models for both the binary and multi-class flare prediction. Models produce a set of probabilities determined based on the prediction mode.

1 adaptive average pool layer, and three fully connected layers. Secondly, we consider a deeper architecture, VGG16 [7], to study whether the performance improves with more layers as it complements the AlexNet model by adding more convolutional layers to the network and using same-sized smaller convolutional kernels of 3×3 for all convolutional layers whereas AlexNet uses variably-sized kernels of 11×11, 5×5 and 3×3. VGG16 consists of 13 convolutional layers, 5 maxpool layers, 1 adaptive average pool layer, and 3 fully connected layers. Finally, we use another landmark CNN model, ResNet34 [8]. It further complements the VGG16 architecture by allowing the network to train deeper layers with less number of parameters. However, it is different from AlexNet and VGG16 in the sense that it takes residuals from each layer and uses them in the subsequent connected layers. ResNet34 has 34 convolutional layers where the first layer has a kernel of 7×7 and the rest have 3×3 kernels with one max pool layer, one adaptive average pool layer and one fully connected layer. The main motivation for selecting these architectures is to understand how the performance changes with different architectures with increasing number of layers and we use the simplest architectures giving consideration to the size of our dataset which is relatively small for deep learning models.

We use all these three models with the transfer learning based approach and exploit the pretrained model weights to improve model training performance in two modes: binary and multi-class where the final layer outputs two and three softmax probabilities respectively. The above architectures trained on binary modes outputs two softmax probabilities for two classes which are then interpreted as no-flare and flare. Similarly, for multi-class modes, the models output three softmax probabilities interpreted as no-flare, mild-flare and strong-flare.

These models are pretrained on the ImageNet dataset [20] which requires a 3-channel image as input to the network. Since the data we use are compressed images of solar magnetograms (which are greyscale), we add a convolutional layer at the beginning of the network which accepts 1-channel input with a 3×3 kernel using size-1 stride, padding and dilation, and outputs a 3-channel image as shown in Fig. 2. This added CNN layer is initialized using Kaiming Initialization in "fan-out" mode [21] for all three models in both binary and multi-class modes. Furthermore, to efficiently exploit the pretrained weights regardless of the architecture of these models, which expects input of different dimensions with 3-channels, we use an adaptive average pooling layer in each models after complete feature extraction using convolutional layer and just before the fully-connected layer to match the dimension on our image input size of 512 × 512.

5 Experimental Evaluation

To train a deep learning model with higher predictive accuracy scores, it is essential to configure the hyperparameters, select an optimization algorithm, and a proper loss function. In addition, it is equally important to prepare the dataset that allows the models to generalize better while training and is sufficiently representative to validate the models. In this section, we elaborate our dataset settings, model implementation, and hyperparameter configurations we have used in this work that directly influence the performance of our models. Furthermore, we present the results of our experiments and the skill scores that characterize the predictive performance of our models in a near-operational setting.

5.1 Experimental Settings

Dataset: In this work, we used compressed images of full-disk line-of-sight magnetograms in bi-daily fashion sampled at 00:00 UT and 12:00 UT for each day. These images are labeled based on maximum GOES peak X-ray flux from 00:00 UT to 24:00 and 12:00 UT to next day 12:00 UT. We ready our cross-validation by dividing our entire data into four time-segmented partitions for both ≥C4.0 and ≥M1.0 class predictions in binary prediction modes and <C4.0, ≥C4.0 to <M1.0, and ≥M1.0 in multi-class modes. Each of these partitions has three months of data from all years included in the entire dataset. The data in Partition-1 contains images from the months of January to March, Partition-2 from April to June, Partition-3 from July to September, and Partition-4 from October to December.

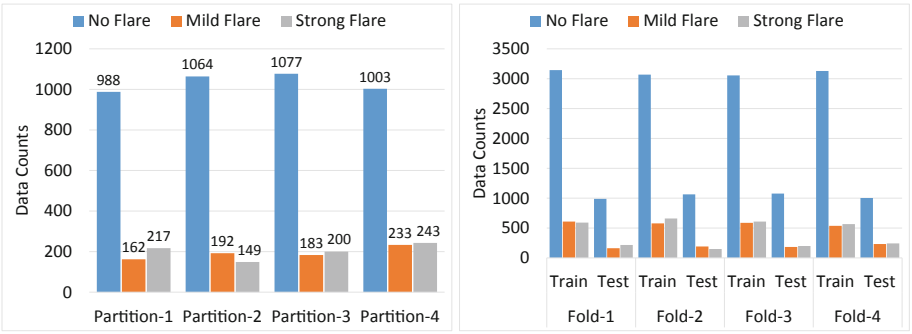
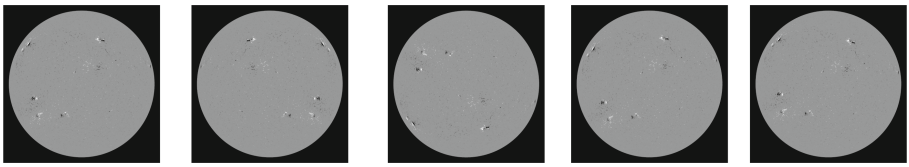


Fig. 3. (a)Time-segmented distribution of data in tri-monthly separated partitions indicating the number of instances for each classes (no-flare, mild-flare and strong-flare) (b) Distribution of 4-fold CV dataset into training and test set created from time-segmented partitions.

Here, this partitioning of the dataset is created by dividing the data timeline from Dec 2010 to Dec 2018 into four partitions on the basis of months rather than chronological partitioning, to incorporate approximately equal distribution of flaring instances in every fold for training and testing the model. As mentioned earlier, we perform two variations of binary predictions: (i) for $\geq$C4.0-class flares, we denote mild-flare and strong-flare as flaring instances and (ii) for $\geq$M1.0 class flares, we exclude the mild-flares, i.e. $\geq$C4.0 to $<$M1.0 from the dataset with a motive to increase the separability in two classes of flares and no-flares. In doing so, all of the four partitions for both the binary prediction modes includes approximately equal number of flare instances. For multi-class prediction mode, we include our entire dataset and the respective partitions contain almost equal number of instances for mild-flare and strong-flare across each partitions as shown in Fig. 3.(a).

We then create the 4-fold CV dataset from the aforementioned partitions where we use three partitions for training the model and the remaining one for testing (validating) the model, ensuring that both the training and test set has data from each year (i.e., Dec 2010 to Dec 2018). First fold (Fold-1) of our 4-fold CV dataset contains data from January to March as the test set and the rest 9 months as training set. Similarly, the second fold (Fold-2) contains the data from April to June as the test set and the rest 9 months of the data as the training set. We use data from July to September and October to December as the test set in the third fold (Fold-3) and fourth fold (Fold-4) with the remaining 9 months of data as the training set respectively. Note that, each fold in 4-fold CV dataset for three different prediction modes: (i) $\geq$C4.0-class binary prediction mode considers mild-flare and strong-flare as flare class (ii) $\geq$M1.0 binary prediction mode do not include the mild-flare into the dataset and denote strong-flares as flare class and (iii) multi-class mode includes the entire dataset as shown in Fig. 3.(b).



(a) Original (b) Horizontal Flip (c) Vertical Flip (d) +5° rotation (e) -5° rotation

Fig. 4. (a) An example compressed magnetogram observed by HMI on 2011–01-01 12:00:00 UT (b) Augmented data sample after applying horizontal flipping. (c) Augmented data sample after applying vertical flipping. (d) Augmented data sample after applying $+5^\circ$ rotation. (e) Augmented data sample after applying -5° rotation.

Implementation: In our experiments, we trained the AlexNet, VGG16 and ResNet34 models with Stochastic Gradient Descent (SGD) as an optimizer and Negative Log-Likelihood (NLL) as our loss function for both binary and multi-class predictions. This implementation of loss function is the generalized version of cross-entropy loss and is not limited to the binary classification problem, hence we use logarithmic-softmax as an activation to the output layer to make it compatible with NLL loss. The NLL loss we use to train our CNN model is:

$$L = -\frac{1}{N}\sum_{i=0}^{N}\sum_{j=0}^{J} y_j \cdot log(\hat{y_j}) + (1 - y_j) \cdot log(1 - \hat{y_j}) \tag{1}$$

where y_j represents the actual j^{th} class label, $\hat{y}$ represents the predicted class label for the j^{th} class and N is the batch-size. For binary predictions, $j = 2$ and for our multi-class prediction mode, $j = 3$. To track the performance improvement of our model, we validate our model with test data in every epoch. An important setting of these experiments is the use of dynamic learning rate which is initialized at 0.01 and reduced by a factor of 10%, if the validation loss do not improve for four consecutive epochs. We use the mini-batch strategy to obtain a faster convergence where the weights are updated after each batch and all of our models are trained up to 80 epochs until weights stability.

We perform four experiments using the 4-fold non-chronological CV dataset and with each architecture in both binary and multi-class prediction modes. Although all of our data partitions have approximately equal numbers of flaring instances, there still exists a prevailing class-imbalance issue. To address the class-imbalance issue, we use data-augmented oversampling; i.e., we oversample the training data after data augmentation only for flaring instances in both binary and multi-class prediction modes so that every batch includes balanced flare and no-flare instances. We use three data-augmentation techniques: vertical flipping, horizontal flipping, and $+5^\circ$ to -5° rotations only on flaring instances included in the training set. Note that the rotations are limited to 5°C as to not impact the preferred locations of active regions (which are limited to activity belts [22]). The Fig. 4 shows the augmented samples of compressed images of magnetograms. These augmented images are then concatenated to the original

training set and then we oversample the flaring instances to create balanced batches for training. Considering the limited amount of data, using oversampling and data augmentation has an advantage that makes the use of entire acquired data, when compared to undersampling [23].

To quantify the performance of our models, we create a classical contingency matrix for both of our binary operating modes, which includes information on True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Note that, in the context of our flare prediction task, flare class in either of the modes is considered as the positive outcome while no-flare is the negative. Using these four outcomes we use two widely used performance metrics in space weather forecasting, True Skill Statistics (TSS, shown in Eq. 2) and Heidke Skill Score (HSS, shown in Eq. 3) to evaluate our models.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \tag{2}$$

$$HSS = 2 \times \frac{TP \times TN - FN \times FP}{((P \times (FN + TN) + (TP + FP) \times N))} \tag{3}$$

Here, $N = TN + FP$ and $P = TP + FN$. Furthermore, for multi-class prediction modes we employ multi-category TSS and HSS as shown in Eq. 4 and 5 respectively [24].

$$TSS = \frac{\frac{1}{N}\sum_{i=1}^{K} n(F_i, O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{K}(N(O_i))^2} \tag{4}$$

$$HSS = \frac{\frac{1}{N}\sum_{i=1}^{K} n(F_i, O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)} \tag{5}$$

where $n(F_i, O_j)$ denotes the number of predictions in category i that had actual observations (ground truth) in category j, $N(F_i)$ denotes the total number of predictions in category i, $N(O_j)$ denotes the total number of observations in category j, and N is the total number of instances in testing set.

TSS values range from -1 to 1, where 1 indicates all correct predictions, -1 represents all incorrect predictions, and 0 represents no-skill, often transpiring as the random or one-sided (all positive/all negative) predictions. It is defined as the difference between True Positive Rate (TPR) and False Positive Rate (FPR). One important characteristic of TSS is that it does not account for class-imbalance ratio in the dataset and hence treats false positives (FP) and false negatives (FN) equally.

Similarly, HSS measures the forecast skill of the models over an imbalance-aware random prediction and it ranges from $-\infty$ to 1, where 1 represents the perfect skill and 0 represents no-skill gain over a random prediction. It is common practice to use HSS for the solar flare prediction models (similar to weather predictions where forecast skill has more value than accuracy or single-class precision), due to the high class-imbalance ratio present in the datasets.

5.2 Evaluation

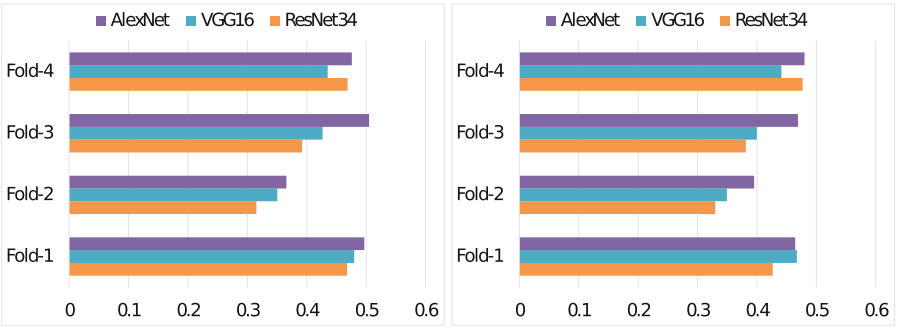
Our flare prediction model is trained as a CNN-based binary classifier where we predict flares in two binary modes with ≥C4.0 and ≥M1.0 as thresholds and a multi-class classifier where we predict flares <C4.0 as no-flares (NF), ≥C4.0 to <M1.0 as mild-flares (MF), and ≥M1.0 as strong-flares (SF). The output of our model is binary (flare/no-flare) predictions and multi-class (no-flare/mild-flare/strong-flare) within the next 24 h. We compare the predictions of our models with maximum GOES peak X-ray flux at 00:00 UT and 12:00 UT with a prediction window of 24 h. We use TSS and HSS metrics to measure the predictive performance of our models.

We summarize the skill scores of all our models in Table 1. The table contains the average skill scores for all three models in binary and multi-class prediction modes with standard deviations across 4-folds computed with confidence level of 95%. These are the stable final epoch cross-validated results obtained by training the models for 80 epochs and validating in every epoch, however, since the ResNet34 model doesn't get fully stable until then, so we compute the average of last five epochs in all prediction modes.

We employ 4-fold cross-validation using the tri-monthly partitioned dataset for evaluating our models as discussed in Sect. 5.1. The TSS and HSS scores obtained from our CV experiments for all three models in binary modes (i) ≥C4.0 class and (ii) ≥M1.0 class are shown in Fig. 5 and Fig. 6, respectively. After training all of our models, we get our best results using the AlexNet model for both binary as well multi-class modes. For binary predictions in ≥C4.0 class modes, all three architectures have relatively low fluctuations with the highest TSS and HSS scores obtained using the AlexNet model. When higher C-class flares filtered from the dataset, we observe an overall increase in both TSS and HSS scores with an exception at Fold-3 results. However, in doing so, the scores have a greater fluctuations across the folds for all the models. While the skill score fluctuations are common in flare prediction studies, Partition-2 includes the most difficult instances to predict, which essentially perturb the overall trend.The best results of our models are comparable to the state of the art deep learning-based flare predictors in the combined performance and and hence provides the evidence that applying a deep learning-based approaches has a high potential for full-disk flare predictions.

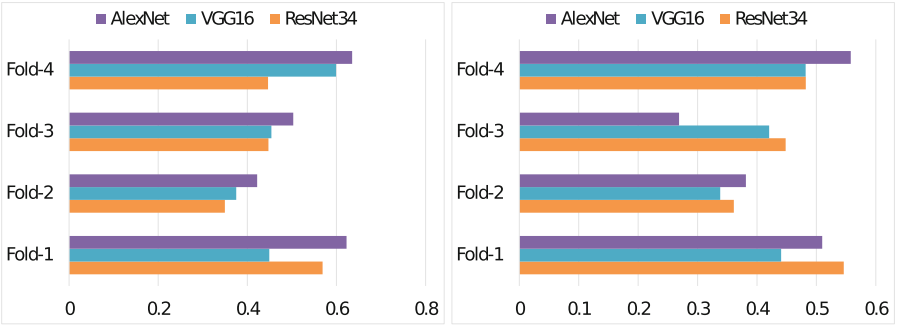
Table 1. Average TSS and HSS skill scores with standard deviation measured at 95% confidence level for all of our models

	Binary (≥C4.0)		Binary (≥M1.0)		Multi-class	
Models	TSS	HSS	TSS	HSS	TSS	HSS
AlexNet	**0.47 ± .06**	**0.46 ± .03**	**0.55 ± .09**	**0.43 ± .11**	**0.36 ± .04**	**0.31 ± .02**
VGG16	0.43 ± .05	0.42 ± .04	0.47 ± .08	0.43 ± .05	0.30 ± .04	0.29 ± .04
ResNet34	0.42 ± .06	0.41 ± .05	0.46 ± .08	**0.46 ± .07**	0.26 ± .05	0.28 ± .05



(a) TSS (b) HSS

Fig. 5. (a) Binary (≥C4.0) prediction performance of our models measured in TSS for each fold in 4-fold CV. (b) Binary (≥C4.0) prediction performance of our models measured in HSS for each fold in 4-fold CV.



(a) TSS (b) HSS

Fig. 6. (a) Binary (≥M1.0) prediction performance of our models measured in TSS for each fold in 4-fold CV. (b) Binary (≥C4.0) prediction performance of our models measured in HSS for each fold in 4-fold CV.

In addition to binary modes, we also evaluated the performance of our trained models in multi-class mode, using multi-class versions of TSS and HSS. Similar to the earlier experiments, AlexNet-based models provided relatively better scores compared to Resnet34 and VGG16 which is presented in Fig. 7 showing the detailed results for each folds. The averaged scores in the last column of Table 1 show that both the skill scores have a relatively low fluctuation ($\sim \pm 0.02$) and our model creates stable predictions for flare prediction. The better predictive performance of AlexNet over other two reasonably advanced models in all of our experiments can be attributed to it's simplicity in the architecture (in terms of number of layers) and the total number of instances in our dataset which is relatively small for deep learning based models.

We also present a set of aggregated contingency tables to better explain the performance of multi-class predictors in Fig. 8. Note that the individual cell

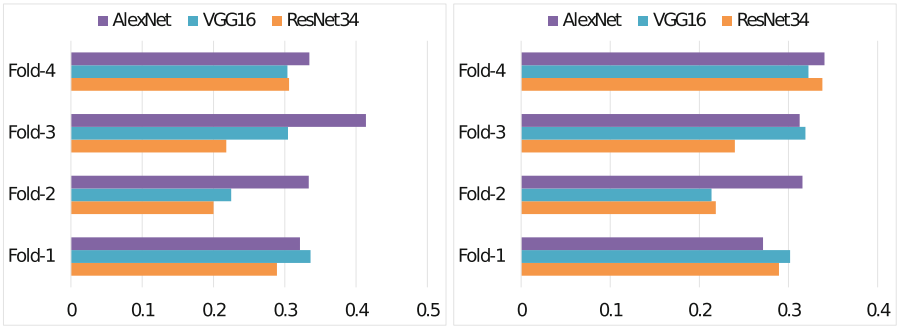


Fig. 7. (a) Multi-class prediction performance of our models measured in TSS for each fold in 4-fold CV. (b) Multi-class prediction performance of our models measured in HSS for each fold in 4-fold CV

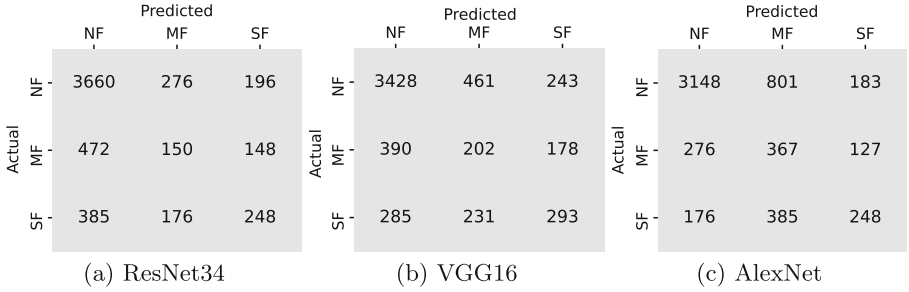


Fig. 8. 4-fold aggregated confusion matrices for multi-class predictions where NF, MF, SF indicates no-flare, mild-flare, and strong-flare respectively for three models (a) ResNet34 (b) VGG16 (c) AlexNet

values are found by summing the values from four contingency matrices obtained in each fold. As expected, multi-class classification is a more difficult prediction problem and the results often show greater shifts between neighboring class label pairs (NF-MF and MF-SF). The aggregated confusion matrix show that, for all three models, mild-flare class has a higher number of false-negatives, which is anticipated, since it lies as the border class between the other two and there is a resemblance of phenomena and therefore a strong likelihood of misclassifying a C5.0-class flare as no-flare ($> C4.0$), or C9.0 flare as strong-flare ($> M1.0$). This is more visible in the aggregated confusion matrix for AlexNet, which is our best model for multi-class predictions, suggests that, the higher number of false-negatives are in the neighboring classes (e.g., NF predicted as MF). Finally, our results show that the predictive performance of AlexNet-based models are satisfactory and can be used under operational settings, first, because it gives more robust results, and second due to its simpler architecture allowing

a computationally efficient platform for near real-time predictions even in the case of large ensembles.

6 Conclusion and Discussion

In this work, we implement CNN-based binary flare prediction models for both ≥C4.0 and ≥M1.0 class prediction modes and one multi-class flare prediction model with three classes: (i) <C4.0 as no-flare class , (ii) ≥C4.0 to <M1.0 as mild-flare class, and (iii) ≥M1.0 as strong-flare class using transfer learning with AlexNet, VGG16 and ResNet34 models. We built efficient flare prediction models having predictive performance comparable to state-of-the-art models using full-disk line-of-sight magnetograms which overcome the prediction ability of active region-based models where the prediction is limited to central locations (within ±70∘). Furthermore, we select a specific threshold of C4.0 for flare prediction since most of the eruptive flares (flares with associated CMEs) are observed to have peak X-ray flux above C4.0 and has an ability to make an impact on near-Earth space in most cases.

For our experiments to make binary predictions in ≥M1.0 class mode, we exclude the data instances indicating the mild-flares to widen the decision boundary for flare and no-flare instances. In order to mitigate the prevailing issue of class-imbalance across our dataset, we use data-augmented oversampling. Based on our experimental results, we observe that AlexNet based model outperforms other two models in both the binary and multi-class prediction modes. This result can be attributed to AlexNet's simple architecture and the total number of images in our dataset which may not be sufficient for depeer models like VGG16 and ResNet34. The results of all cross-validated experiments suggests that the AlexNet model can be used in an operational setting to perform near-real time flare predictions. To reproduce this work, the source code and detailed experimental results can be accessed from our open source repository [25].

It is also important to mention that our models use point-in-time observations of magnetogram images and do not identify the active regions contributing to the corresponding flaring event. Furthermore, the Eastern limb flares (whose active regions only become visible after predictions are issued) limit the prediction capabilities of our models due to unavailable active region information. Subsequently, we intend to use other deeper variants of CNN-based architectures along with integration of different dimensions of solar data products such as vector magnetograms, intensitygrams, dopplergrams, and extreme ultraviolet images. One important aspect of this work is the utilization of a more practical threshold of C4.0 in flare forecasting which is discussed in literature but not considered in practical implementation, therefore, we will continue our experiments by optimizing these thresholds for different architectures and modes to further improve the flare prediction models. To improve the performance of our current models we also aim to build hybrid models by combining active region-based counterparts to obtain more robust ensemble flare prediction models.

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

References

1. Shea, M., Smart, D., McCracken, K., Dreschhoff, G., Spence, H.: Solar proton events for 450 years: the carrington event in perspective. Adv. Space Res. **38**(2), 232–238 (2006)
2. 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. Astrophys. J. **869**(2), 91 (2018)
3. Li, X., Zheng, Y., Wang, X., Wang, L.: Predicting solar flares using a novel deep convolutional neural network. Astrophys. J. **891**(1), 10 (2020)
4. Fletcher, L., et al.: An observational overview of solar flares. Space Sci. Rev. **159**(1–4), 19–106 (2011)
5. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
6. Krizhevsky, A.: One weird trick for parallelizing convolutional neural networks (2014)
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
9. Matsugu, M., Mori, K., Mitari, Y., Kaneda, Y.: Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw. **16**(5–6), 555–559 (2003)
10. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Ishii, M.: Deep flare net (DeFN) model for solar flare prediction. Astrophys. J. **858**(2), 113 (2018)
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. Astrophys. J. **856**(1), 7 (2018)
12. Szegedy, C., et al.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2015
13. 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, July 2017
14. Bobra, M.G., et al.: The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: SHARPs – space-weather HMI active region patches. Solar Phys. **289**(9), 3549–3578 (2014)
15. Ahmadzadeh, A., Aydin, B., Georgoulis, M.K., Kempton, D.J., Mahajan, S.S., Angryk, R.A.: How to train your flare prediction model: revisiting robust sampling of rare events. Astrophys. J. Supplement Ser. **254**(2), 23 (2021)
16. Muller, D., et al.: JHelioviewer: visualizing large sets of solar images using JPEG 2000. Comput. Sci. Eng. **11**(5), 38–47 (2009)

17. Alzubaidi, L., et al.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data **8**(1), 1–74 (2021). https://doi.org/10.1186/s40537-021-00444-8
18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
19. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). IEEE, August 2017
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, June 2009
21. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification (2015)
22. Pulkkinen, P.J., Brooke, J., Pelt, J., Tuominen, I.: Long-term variation of sunspot latitudes. Astronomy Astrophys. **341**, L43–L46 (1999)
23. 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 International Conference on Big Data (Big Data). IEEE, December 2019
24. WWRP/WGNE Joint Working Group on Forecast Verification Research: Forecast verification issues, methods and FAQ, January 2015. https://www.cawcr.gov.au/projects/verification/
25. DMLab: Source Code. https://bitbucket.org/gsudmlab/fulldisk_simbig/src/master