Probalistic Prediction of Dst Storms One-Day-Ahead Using Full-Disk SoHO Images

A. Hu1,2, C. Shneider1, A. Tiwari1, and E. Camporeale2,3

1Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, 2CIRES, University of Colorado, Boulder, CO, USA,
3NOAA Space Weather Prediction Center, Boulder, CO, USA

Abstract  We present a new model for the probability that the disturbance storm time (Dst) index exceeds −100 nT, with a lead time between 1 and 3 days. Dst provides essential information about the strength of the ring current around the Earth caused by the protons and electrons from the solar wind, and it is routinely used as a proxy for geomagnetic storms. The model is developed using an ensemble of Convolutional Neural Networks that are trained using Solar and Heliospheric Observatory (SoHO) images (Michelson Doppler Imager, Extreme ultraviolet Imaging Telescope, and Large Angle and Spectrometric Coronagraph). The relationship between the SoHO images and the solar wind has been investigated by many researchers, but these studies have not explicitly considered using SoHO images to predict the Dst index. This work presents a novel methodology to train the individual models and to learn the optimal ensemble weights iteratively, by using a customized class-balanced mean square error (CB-MSE) loss function tied to a least-squares based ensemble. The proposed model can predict the probability that Dst < −100 nT 24 hr ahead with a True Skill Statistic (TSS) of 0.62 and Matthews Correlation Coefficient (MCC) of 0.37. The weighted TSS and MCC is 0.68 and 0.47, respectively. An additional validation during non-Earth-directed CME periods is also conducted which yields a good TSS and MCC score.

Plain Language Summary  Geomagnetic storms pose one of the most severe space weather risks to our space-borne and ground-based electronic instruments, such as GNSS and radio transmission systems. Dst is one of the most accurate geomagnetic storm indicators. Hence, those storm can be predictable if Dst can be forecasted. Currently, the best Dst model can only predict Dst in several hours. In this study, we present a machine learning based ensemble method to predict the Dst 1–3 days in advance from solar images.

1. Introduction

Geomagnetic storms pose one of the most severe space weather risks to our space-borne and ground-based electronic instruments, such as GNSS and radio transmission systems. A geomagnetic storm can be indicated by several geomagnetic indices such as Kp, ap, and the disturbance storm time (Dst) index (Rostoker, 1972). These indices are related to the perturbation of the geomagnetic field as measured on local regions on Earth at middle, high, and low latitudes, respectively. Although it is now recognized that a single index is not able to capture and define all geospace storms (Borovsky & Shprits, 2017), they are routinely used by space weather operational agencies as proxies for geomagnetic activity (see, e.g., https://www.swpc.noaa.gov/products/geospace-geomagnetic-activity-plot). Here, we focus specifically on Dst, given the large amount of literature devoted to its prediction, notably using data-driven and machine learning techniques (Camporeale, 2019). Dst is understood to be a proxy for ring current density (Liemohn et al., 2001) and it is currently defined by using quasi real-time geomagnetic field measurements from four equatorial ground magnetometer stations: Hermanus, Honolulu, San Juan and Kakioka (Sugiura & Kamei, 1991).

Most of the current models predict Dst based on solar wind parameters such as the North-South component of the interplanetary magnetic field (IMF) Bz (Saiz et al., 2008). Neural networks have been widely used in modeling Dst empirically. Lundstedt et al. (2002) was one of the first to implement a multi-layer perception (MLP) neural network based on IMF Bz, solar wind density and velocity, in order to forecast Dst 1-hr in advance. Saiz et al. (2008), Bala and Reiff (2012), and Lazzús et al. (2017) presented models to forecast Dst up to 6 hr in advance. A Gaussian Process model has been introduced by Chandorkar et al. (2017) and Chandorkar and Camporeale (2018) and later combined with a a long short-term memory (LSTM) architecture in Gruet et al. (2018) to provide a probabilistic forecast up to 6 hr in advance. An ensemble learning algorithm has been used in Xu
Figure 1. Time history of Dst during 1996–2010. x axis is date and Y axis is Dst value. The orange crosses denote peak values smaller than −100 nT, used for defining storm events considered in this study.

Figure 2. An example of the selection criterion to define the time range for one storm event. The Dst peak occurs on 23 October 1996. The nearest positive Dst values before and after the peak occur on 18 October and 03 November, respectively. The whole storm range is defined between 17 October 1996 and 04 November 1996 with a 24-hr buffer zone. The list of selected storm events can be found in Table 1.
et al. (2020), Laperre et al. (2020) evaluated the performance of a LSTM model based on a Dynamical Time Warping (DTW) method.

All the empirical models mentioned above are trained from solar wind data which are measured in quasi real-time by the ACE or DSCOVR satellites orbiting around the first Lagrangian point (L1), or alternatively using the NASA OMNI database (https://omniweb.gsfc.nasa.gov/). Hence, in an operational setting, their lead-time would be limited to only a few-hours ahead.

In this paper, we aim to predict $Dst$ with a longer lead time (in the range of 1–3 days ahead) using solar images from the Solar and Heliospheric Observatory (SoHO) as inputs. SoHO is a joint mission between the National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA) and was the first space-based telescope to serve as an early warning system for space weather. Solar images can be observed by a suite of on-board instruments on SoHO (Domingo et al., 1995), including the Michelson Doppler Imager (Scherrer et al., 1995) (MDI) for the solar photosphere, the Extreme ultraviolet Imaging Telescope (Delaboudinière et al., 1995) (EIT) for the stellar atmosphere to low corona, and the Large Angle and Spectroscopic Coronagraph (Brueckner et al., 1995) (LASCO). Although used much less than solar wind data for forecasting purposes, it is known that a significant correlation exists between EIT and $Dst$ index. A semi-physical model called “Anemomilos” is then developed based on this relationship to predict $Dst$ index in 6 days (Tobiska et al., 2013). This model became part of the US Space Force HASDM predictions in 2012. Upendran et al. (2020) has pointed out that the correlations between solar images and solar wind parameters are most significant during the fast solar wind.

Obviously, by setting the problem as a 1–3 days ahead forecast, we have to accept that we cannot achieve the accuracy seen in few 1–6 hr ahead forecast models, that currently report Root Mean Square Errors of the order of 10 nT or less. Therefore, as a first step, we set the problem as a classification task, aiming at forecasting the probability that $Dst$ exceeds a certain threshold (hereinafter referred to as “$Dst$ probability”). In this study, we focus on strong storms having a $Dst$ threshold of $-100$ nT, and aiming at producing a probabilistic forecast 1 day ahead of a given solar image.

In addition, although most operational applications require a deterministic $Dst$ value, the predicted probabilistic $Dst$ forecast could also be used to improve current space weather models, for example, for running ensemble simulations of mass density forecasting, which is one of the top priority for the predictability of low-Earth-orbit (LEO) satellite trajectories (Licata et al., 2020). $Dst$ plays a major role in mass density modeling such as Jacchia-Bowman2008 (JB2008) (Bowman et al., 2008). Hence, The forecast $Dst$ probability would be helpful to assess the uncertainty of those models.

We train a machine learning (ML) technique called convolutional neural network (CNN) to forecast the probability that $Dst$ exceeds the pre-defined $-100$ nT threshold from 1 to 3 days in advance (i.e., the prediction is in the form of a time series of probabilities). CNN has been recently used in space weather applications, for example, by Siciliano et al. (2021), Upendran et al. (2020), Li et al. (2020), Ruwali et al. (2020), and Park et al. (2018). By using the presented technique in an operational setting, a forecaster would have access to several predictions issued with different lead-times. Hence, we face a classical problem in ensemble learning, namely how to combine different predictions by applying different weights to different lead-times.
In this work, we restrict to a static weighting scheme (i.e., the weights are learned on a training set and do not change with different inputs or solar wind conditions), opposite to dynamic weights (Polikar, 2012). We solve the ensemble problem by introducing a new, customized, complementary cumulative distribution function (CCCDF) based least-squares (LS) method to find the optimal weights.

The paper is divided as follows. Section 2 introduces the data used for this study, the criterion to select storm times and the corresponding time periods covered. Section 3 describes the methodology, including the designed machine learning architecture, the optimization method, and the performance metrics for assessment. Section 4 presents the results of the developed model, and emphasizes the probabilistic nature of the forecast. Finally, in Section 5, we draw conclusions and make final remarks about future directions.

### 2. Data

#### 2.1. Disturbance Storm Time (Dst) Index

The $\text{Dst}$ index is available at 1-hr cadence from the NASA OMNI database. Figure 1 displays the $\text{Dst}$ index in the period 1996–2010. The model is trained, validated and tested on storm events with a $\text{Dst}$ peak smaller than $-100 \text{ nT}$, shown by orange crosses. Overall, 51 such storm periods are selected for this study. In order to define a storm period, we look for the the nearest positive $\text{Dst}$ values immediately before and after each peak, and then extend the time window by a 24-hr buffer zone to make sure that the pre-storm period and the recovery phase are fully included. An example is shown in Figure 2, where the $\text{Dst}$ peak is observed on 23 October 1996. The storm period is defined as ranging between 17 October 1996 and 04 November 1996. With this procedure we make sure that the time intervals are selected in such a way that the negative $\text{Dst}$ peaks do not always occur at the same time within the chosen storm-time window, hence the neural network does not simply memorize. The average period of selected storm events is approximately 15 days. All selected storms, sorted by peak $\text{Dst}$, are listed in Table 1.

As mentioned in Section 1, we would like to solve this classification task by a regression model. Hence, instead of a binary label set (positive/negative), a customized complementary cumulative probability distribution function (CCCDF) of $\text{Dst}$ is used as a target for the CNN model. The CCCDF is shown in Figure 3, and explained briefly below. A cumulative distribution function (CDF) is defined as the integral of a probability density function (PDF) from negative infinity to $x$ and the complementary CDF (CCDF) is $1 - \text{CDF}$ or the integral from positive infinity to $x$. The CDF($x$) is the probability that a random variable has a value less than $x$. Conversely, the CCDF($x$) gives the probability that the variable under consideration is larger than $x$. The customized CCDF (CCCDF) for $\text{Dst} = -100 \text{ nT}$ is defined as in Equations 1 and 2. Note that $\text{CCCDF}(\text{Dst} = -100) = 0.5$ by construction.

$$
\text{CCCDF}(x) = \frac{(\text{CDF}(x) - \text{CDF}(-100))}{1 - \text{CCDF}(-100)} + 0.5 \text{ for } x \leq -100
$$

$$
\text{CCCDF}(x) = \frac{(\text{CDF}(x) - \text{CDF}(-100))}{\text{CCDF}(-100)} + 0.5 \text{ for } x > -100
$$

#### 2.2. SoHO Mission

The two-hourly SoHO data sets used in this work for the period 01 May 1996 to 12 April 2011 are derived from the following public domain resources: the NASA Solar Data Analysis Center's (SDAC) Virtual Solar Observatory (VSO) (https://sdac.virtualsolar.org/cgi/search) and Stanford University's Joint Science Operation Center (JSOC) http://jsoc.stanford.edu/MDI/MDI_Magnetograms.html. All SoHO products and their details are shown in Table 2.

More than 20,000 SOHO images can be provided from three on-board instruments, including the Michelson Doppler Imager (MDI) for the solar photosphere, the Extreme ultraviolet Imaging Telescope (EIT) for the stellar atmosphere to low corona, and the Large Angle and Spectrometric Coronagraph (LASCO) covering the corona

| No. | Start time     | End time     | Min. $\text{Dst}$ (nT) |
|-----|----------------|--------------|------------------------|
| 44  | 2002-12-17 08:00:00 | 2002-12-24 16:00:00 | −75 |
| 45  | 2005-10-29 09:00:00 | 2005-11-04 18:00:00 | −74 |
| 46  | 2003-01-27 13:00:00 | 2003-02-15 13:00:00 | −74 |
| 47  | 2007-03-21 10:00:00 | 2007-03-27 00:00:00 | −72 |
| 48  | 2002-02-26 17:00:00 | 2002-03-05 01:00:00 | −71 |
| 49  | 2010-04-30 11:00:00 | 2010-05-12 17:00:00 | −71 |
| 50  | 1999-10-08 01:00:00 | 1999-10-22 04:00:00 | −67 |
| 51  | 2003-02-24 22:00:00 | 2003-03-14 15:00:00 | −67 |
from 1.5 to 30 \( R_s \). Those data have fully covered Solar Cycle 23 and 24. Among them, MDI, EIT with a wavelength of 195 (EIT-195) and LASCO-C2 are used as the inputs of this study.

However, the SDAC data is highly heterogeneous. Not only are there intrinsic differences among these SoHO products (e.g., individual cadence for each channel shown in Table 2), but there is also an irregular assortment of image file sizes and processing levels. All products require calibration before they can be used for the neural networks. In addition, each channel needs to be synchronized with a fixed cadence (i.e., 2 hr in this study). An example of the calibrated data are shown in Figure 4.

A pipeline has been created and published by Shneider et al. (2021) for automatically downloading, cleaning and synchronizing these original images from SDAC and VSO. A machine-learning-ready image data set is then provided which is a valuable resource for the space weather community.

3. Methodology

The goal of this study is to estimate the \( Dst \) probability 1–3 days ahead of the time when full-disk SoHO images are taken. For the sake of clarity we discuss here the algorithm for 1-day ahead prediction, with the understanding that all times are correspondingly shifted for 2 and 3 days ahead predictions.

3.1. Customized Class-Balanced Convolutional Neural Networks (CB-CNN)

Several machine learning approaches can be used for a probabilistic prediction task. We have compared naive Bayes, MLP and CNN. Among them, CNN has yielded the most reliable and robust performance (results are not shown here).

CNN is a commonly used neural network architecture, widely used in computer vision (Gu et al., 2018), in solar image processing (Baso & Ramos, 2018; Dos Santos et al., 2021; Illarionov & Tlatov, 2018; Upendran et al., 2020), and recently in plasma and space physics applications (Hu et al., 2020; Siciliano et al., 2021). Hence, in this study, we have opted to use a customized class-balanced CNN. This is because of the imbalance between the number of positive samples (\( Dst \leq -100 \) nT) and the number of negative samples (\( Dst > -100 \) nT), which is approximately 10% of all training samples. As a result of using a class-balanced target, the CB-CNN model performs better than a vanilla CNN model, when the target \( Dst \) is near the classification threshold. A brief introduction of the CNN architecture and the optimization methods used in the training is listed in Table 3.

### Table 2

| Instrument | Detector | Observed region | \( \lambda \) (Å) | Cadence (min) | Date range             |
|------------|-----------|-----------------|------------------|---------------|------------------------|
| MDI        | MDI       | Full disk       | 6768 (Ni I)      | \~96          | 1996.05.01–2011.04.12  |
| EIT        | EIT       | Full disk       | 171 (Fe IX/X)    | \~360         | 1996.01.01 →           |
| EIT        | EIT       | Full disk       | 195 (Fe XII)     | \~12          | 1996.01.01 →           |
| EIT        | EIT       | Full disk       | 284 (Fe XV)      | \~360         | 1996.01.01 →           |
| EIT        | EIT       | Full disk       | 304 (He II)      | \~360         | 1996.01.01 →           |
| LASCO      | C2        | Corona (1.5–6 \( R_s \)) | Visible | \~20          | 1995.12.08 →           |
| LASCO      | C3        | Corona (3.5–30 \( R_s \)) | Visible | \~20          | 1995.12.08 →           |

Note. \( \lambda \) (Å) is wavelength measured in angstroms, and \( R_s \) is the Sun’s radius. LASCO C1 (1.1–3 \( R_s \)) is not included in this work since it was only operational till 9 August 2000. MDI, Michelson Doppler Imager; EIT, Extreme ultraviolet Imaging Telescope; LASCO, Large Angle and Spectrometric Coronagraph.
By denoting with $T$ the time at which the input images are taken, a multi-target CNN model is first trained to predict the $Dst$ probability in the time range $[T + 26, T + 48]$, with a time resolution of 2 hr. That is, 12 probability values are output for each input. This means that, at any given time, we have 12 probabilities that have been predicted between 26 and 48 hr ahead. Note that, due to the variability of the Sun and because we are using full disk images, predictions with shorter time lags are not necessarily more accurate than ones with longer time lags. Because we eventually want to merge those predictions (with different time lags) into a unique, reliable prediction, we do not use a standard loss function for binary classification (such as, e.g., binary cross-entropy), but a customized, class-balanced, mean square error.

In our application, a class-balanced loss function developed by Cui et al. (2019) is used to deal with the large imbalance between positive and negative labels. In addition to that, we want to penalize more the incorrect predictions that are closest to the decision boundary $CCCDF = 0.5$ (corresponding to the threshold $Dst = -100$). Hence, a customized weight is designed to artificially increase the cost function for the possibly ambiguous samples with $Dst$ near $-100$ nT. The cost function is defined as:

![Figure 4. Examples of various Solar and Heliospheric Observatory (SoHO) products (a–c). Colors similar to those used on the NASA SoHO site have been used.](image-url)
Table 3  
Parameter Selection of CNN and the Corresponding Optimization Method Used in Training

| Parameter                  | Value               |
|----------------------------|---------------------|
| Input size                 | [3, 256, 256]       |
| Output size                | 12                  |
| Layers                     | 3                   |
| Kernel size                | [3, 5, 5]           |
| Padding mode               | Same                |
| Activation function        | [ReLU, ReLU, ReLU]   |
| Max epochs                 | 100                 |
| Optimization               | AdamW               |
| Learning rate              | 0.0001              |
| Regularization             | Elastic             |
| Cost function              | CB-MSE              |

Cost Function = \[
\sum_{h=26}^{48} w_h \left( \sum_{i=1}^{n_{pos}} E_h^i (1 - \beta) + \sum_{i=1}^{n_{neg}} E_h^i (1 - \beta) \right) \]  

where \( E_h^i = (\hat{P}_h^i - P_h^i) \cos (P_h^i - 0.5) \times 0.9 \times \pi^2 \)  

\( \hat{P}_h^i \) is the model prediction for \( i \)th sample with a delay hour \( h \) (delay hours \( h \) ranging 26, 28, ..., 48 hr), \( P_h^i \) is the corresponding target, \( n_{pos} \) and \( n_{neg} \) are the number of positive and negative samples in a batch, respectively, and \( w_h \) denotes the weight associated to each target. The detailed procedure of optimizing these weights will be further introduced in Section 3.3. \( E \) is a customized square error that penalizes samples whose Dst is near \(-100\) nT, as shown in Figure 5. \( \beta \) is a constant term manually set to 0.9999 according to Cui et al. (2019).

The multi-target CNN trained using the cost function described above, can then be used to forecast the Dst probability from 26 to 48 hr ahead, at a 2 hr rate. Hence, at any given time we have \((48 - 26)/2 + 1 = 12\) different predictions (issued between 26 and 48 hr prior). A natural question then arises on whether one could combine these 12 predictions in order to achieve a prediction that is more accurate than any individual one. In order to do that, we take a weighted average of the 12 predictions, and we estimate the optimal (static) weights by solving a least squares (LS) problem. The details of the weight estimate procedure are shown in Figure 8. Figure 9 shows the learned 12 weights optimized with this procedure, along with their uncertainty (see the leave-one-out procedure described below).

Finally, we notice that the predicted Dst probabilities might not be well-calibrated (i.e., statistically consistent with observations) and that the optimal threshold for binary classification metrics might be different than the standard 50% probability. Hence, the threshold of probability used for metrics (see Section 3.2) is re-calibrated by using a receiver operating characteristic (ROC) curve. ROC curve is an important diagnostic for a probabilistic model that can be used to determine an optimal threshold to separate positives from negatives based on probabilistic predictions. A detailed description of ROC curve can be found in Camporeale et al. (2020). An example ROC curve for the developed model with the CNN is shown in Figure 6. Horizontal and vertical axes denote false positive rate (FPR) and true positive rate (TPR), respectively. The dashed orange line shows TPR equals to FPR (i.e., no skill), while the blue line represents the ROC curve, obtained by defining positives and negatives by progressively changing the probability threshold from 0% to 100%. The red dot represents the optimal/largest value of True Skill Statistics, defined as the difference (TPR-FPR).

Figure 5. Contour map of Equation 4. Left panel is absolute error between \( \hat{P} \) and \( P \); middle panel is the weight in \( E \), which is the second term in Equation 4; and the last panel is final \( E \).
3.2. Metrics

In order to precisely assess the accuracy of a model, it is important that the performance metrics are computed on a test set independent from the training set (so-called hold-out data), hence making sure that the machine learning algorithm does actually learn meaningful patterns and does not merely memorize the training data. A “Leave one out” technique is adopted here. That is a $K$-fold cross validation taken to its logical extreme, with $K$ equal to $N$, the number of selected storm cases. That means that the proposed model is trained on all the data except for one storm window and a prediction is made for that left-out storm. The procedure is repeated $N$ times. Finally, the metrics are computed as averages over the $N$ models. In this study, the top 51 storm windows in the period 01 January 1999 to 10 April 2011 constitute each a fold. The probabilistic predictions can be transformed to binary labels upon defining a probability threshold. In this way we can use standard metrics for binary classification such as the True Skill Statistic (TSS) and Matthews Correlation Coefficient (MCC) (Camporeale et al., 2020):

$$\text{TSS} = \frac{TP}{TP + FN} - \frac{FP}{FP + TN},$$

(5)

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (FN + TN) \times (FP + TN) \times (TP + FN)}},$$

(6)

where TP, FP, TN and FN denotes true positive, false positive, true negative, and false negative numbers respectively. The MCC score is a reliable statistical rate that produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (TP, FP, TN, and FN), proportionally both to the size of positive elements and the size of negative elements in the data set (Baldi et al., 2000). TSS is a useful metric that combines both types of information and should be as close as possible to 1. Those metrics have shown some advantages over the F1 score and accuracy in binary classification evaluation (Chicco & Jurman, 2020). Moreover, an innovative way to evaluate the model accuracy has been developed by Guastavino et al. (2021). This method assign different weights to FPs that anticipate the occurrence of an actual positive event (i.e., “almost hit”). The value-weighted MCC and TSS have been proved more appropriate for decision making processes. Hence, these weighted scores are also considered to assess the model accuracy.
3.3. Ensemble Method

After the CNN model is developed, 12 probabilities can be predicted from the model and each SOHO image set. Although each prediction is per se valid, we have verified that combining those predictions yields a model that outperforms a single individual prediction. Here, we describe the ensemble method that, for simplicity, has been chosen to be a simple linear combination of the 12 probabilities so that the final probability is defined as

$$ A_{ensemble} = \sum_{i=26}^{48} w_i p_i $$

where $p_i$ is the probability of $Dst$ exceeding the −100 nT threshold at time $i$. The timeline of predictions is depicted in Figure 7. Each horizontal bar displays the timeline during an hypothetical event. Yellow blocks denote when Solar and Heliospheric Observatory (SoHO) images are taken. Green blocks are the 12 $Dst$ probability predictions as introduced from Section 3.1. The weights of those 12 predictions from each sample are consistent. From top to bottom, there are $n$ samples during this event. The time shift between nearby samples is 2 hr. All predicted probabilities in one red frame is considered as one probability cluster (i.e., a 12 × 1 vector) for a certain time epoch.

Figure 7. Illustration of the ensemble procedure. Horizontal bar displays the timeline during the whole event. Yellow blocks denote when Solar and Heliospheric Observatory (SoHO) images are taken. Green blocks are the 12 $Dst$ probability predictions as introduced from Section 3.1. The weights of those 12 predictions from each sample are consistent. From top to bottom, there are $n$ samples during this event. The time shift between nearby samples is 2 hr. All predicted probabilities in one red frame is considered as one probability cluster (i.e., a 12 × 1 vector) for a certain time epoch.
Figure 8. Flowchart of modeling procedures.
\[ L = W \cdot P, \]  

(7)

where \( P \) is the design matrix as introduced in Figure 7 (each column of \( P \) contains 12 different probabilities collected with different time lags):

\[
P = \begin{bmatrix}
p_{26h}^{12} & p_{26h}^{13} & \cdots & p_{26h}^{26} \\
p_{28h}^{11} & p_{28h}^{12} & \cdots & p_{28h}^{26} \\
\vdots & \vdots & \ddots & \vdots \\
p_{48h}^1 & p_{48h}^2 & \cdots & p_{48h}^{26}
\end{bmatrix}
\]

(8)

\( L \) denotes the ground truth (i.e., the CCCDF of \( Dst \)), and \( W \) is the sought after weight vector. \( W \) is initially set to a constant vector, and will be optimized by a least-square (LS) method:

\[
W = [w_{26h} \ w_{28h} \ \cdots \ w_{48h}]
\]

(9)

**Table 4**

Accuracy of the Proposed Ensemble Model With LS, SVR and Constant Weights, Together With Single-Target CNN Model

| Ensemble method | TSS  | MCC  | TP   | FP   | TN   | FN   |
|-----------------|------|------|------|------|------|------|
| LS              | 0.62 | 0.37 | 57   | 209  | 1738 | 21   |
| SVR             | 0.15 | 0.18 | 13   | 37   | 1910 | 65   |
| Constant        | 0.32 | 0.23 | 31   | 135  | 1812 | 47   |
| Single          | 0.28 | 0.17 | 103  | 544  | 2097 | 111  |

TSS, True skill statistic; MCC, Matthews Correlation Coefficient; TP, true positive; FP, false positive; TN, true negative; FN, false negative numbers.

**Figure 9.** Boxplot of weights of 12 ensemble predictions. Green lines are median value for each weight, and black dots are anomalies. The upper and lower boundary of blue lines are the maximum and minimum. The upper and lower boundary of blue lines are the first and third quartile within 51 storm events.
In this study, we develop a customized Elastic-net-aided LS method to optimize the weights \( W \) based on the elastic-net regularization scheme proposed by Zou and Hastie (Zou & Hastie, 2005). The residuals are calculated as:

\[
V = L - P \cdot W
\]  

The cost function, or so-called normal function in LS, is defined as:

\[
F = V^T Q V + r_1 \sum W + r_2 \sum W^2
\]  

The first term in \( F \) is a classic weighted cost function where \( V^T \) is the transpose matrix of \( V \). The second and third terms are Elastic net regularization factors that linearly combines the L1 and L2 penalties of the lasso and ridge methods. Based on the experiments on the storms, here we set \( r_1 = 0.05 \), and \( r_2 = 0.95 \). Finally, \( Q \) denotes the “weight,” that is, diagonal matrix with positive and negative samples as defined in Equation 12.

\[
q_{\text{pos}} = \frac{n_{\text{all}}}{n_{\text{pos}}} \cdot q_{\text{neg}} = \frac{n_{\text{all}}}{n_{\text{neg}}}
\]  

The function \( F \) reaches a minimum when the partial derivative of \( F \) with respect to \( W \) equals zero:

\[
\frac{\partial F}{\partial W} = -2V^T Q P + \tilde{r}_1 + 2\tilde{r}_2 W = 0,
\]  

where \( \tilde{r}_1 \) is a 12 × 1 vector and \( \tilde{r}_2 \) is a 12 × 12 unit vector times scale \( r_2 \).

or

\[
P^T Q V = \frac{1}{2} \tilde{r}_1 + \tilde{r}_2 W
\]  

Multiplying \( P^T Q \) with Equation 10, one has

\[
P^T Q V = P^T Q L - P^T Q P \cdot W
\]  

Setting Equation 14 into Equation 15, one has

\[
\frac{1}{2} \tilde{r}_1 + \tilde{r}_2 W = P^T Q L - P^T Q P \cdot W
\]  

Finally,

\[
W = (P^T Q P + \tilde{r}_2)^{-1} \left( P^T Q L - \frac{1}{2} \tilde{r}_1 \right)
\]

### 4. Results

In this section we show the results of our model in terms of the metrics TSS and MCC scores discussed in Section 3.2. It should be noted that, by using the leave-one-out technique, all metrics in this section are calculated based on the combination of all entries in the confusion matrix. The proposed ensemble approach is compared against three alternative approaches: support vector regression (SVR) which is a non-linear ensemble method (Awad & Khanna, 2015), a method where ensemble members are simply averaged (equal weight, denoted as “Constant” in Table 4), and a single individual prediction 24 hr ahead (no ensemble, so-called “Single” in Table 4). Table 4 shows the TSS and MCC for the four methods. One can notice that the LS ensemble method significantly outperforms the single individual prediction model. Moreover, although SVR yields a large TNR, the FPR is also large, resulting in low values for both TSS and MCC.

### Table 5

| Dst threshold | TSS | MCC | TP | FP | TN | FN |
|---------------|-----|-----|----|----|----|----|
| −50 nT        | 0.29| 0.26| 215| 350| 1254| 206|
| −100 nT       | **0.62**| **0.37**| 57 | 209| 1738| 21 |
| −200 nT       | 0.56| 0.14| 3  | 80 | 1940| 2  |

TSS, True skill statistic; MCC, Matthews Correlation Coefficient; TP, true positive; FP, false positive; TN, true negative; FN, false negative numbers.

### Table 6

| Forecast (Days) | TSS | MCC | TP | FP | TN | FN |
|-----------------|-----|-----|----|----|----|----|
| 1               | **0.62**| **0.37**| 57 | 209| 1738| 21 |
| 2               | 0.34| 0.30| 40 | 93 | 1807| 63 |
| 3               | 0.27| 0.17| 44 | 248| 1578| 65 |

TSS, True skill statistic; MCC, Matthews Correlation Coefficient; TP, true positive; FP, false positive; TN, true negative; FN, false negative numbers.
As anticipated in Section 1, the goal of this work is not just to provide a binary classification, but rather to estimate the probability of exceeding predefined thresholds. The LS-ensemble method has been trained and tested for various Dst thresholds (Table 5), different forecast duration (see Table 6) and larger time resolution (Table 7). Table 5 demonstrates that the model performs best when the Dst threshold is $-100$ nT. Corresponding TSS and MCC are 0.62 and 0.37, respectively. Both TSS and MCC decrease when the threshold is set to $-50$ nT. Although TSS performs well with a threshold of $-200$ nT, the corresponding MCC decreases significantly because of the imbalanced labels since very few “very strong” storms ($\leq -200$ nT) occurred during 1999–2009. This implies that the proposed method may identify strong storms ($\leq -100$ nT) better than mild storms ($\leq -50$ nT). Table 6 shows how the performance of the model gets worse with a longer lead-time. From Table 7, we can see that the model trained from 2-hr samples outperforms the model trained from 6-hr samples. This is because a 6-hr samples training set is composed of fewer samples overall. Therefore, the model could be less robust with a larger cadence. Finally, we show in Table 8 the modified TSS and MCC scores proposed in Guastavino et al. (2021) that further improve the accuracy of the method.

A statistic analysis of weights from all of the 51 sub-models is plotted in Figure 9. Green horizontal lines denote the mean weights from all 51 sub-models. Blue bars represent the uncertainty range between the first and the third quartile of the distributions, and black dots are distribution outliers. Figure 9 shows that the predictions at 24–38 delay hours have the largest contributions to the final probability. The weights decrease with the increase of delay hours. It is interesting that the contribution of predictions during 42–46 delay hours are essentially negligible. The error bars and presence of outliers also imply that the weights vary slightly according to different storms. This may be improved by having more representative storm events.

### 4.1. Storm Case Study

In this section we would like to investigate several typical storm cases.

The Halloween storm, caused by a CME, from 25 October 2003 to 05 November 2003 is selected for a case study. This is the biggest storm in the past 20 years. Figure 10 displays the 24-hr ahead Dst probabilities of the developed model (green line, left vertical axis) and the corresponding Dst (blue line, right vertical axis) during the Halloween storm. Black cross and dotted lines are the threshold of probability and Dst index, respectively.

The predicted probabilities can be converted to a binary format with an optimal threshold re-calibrated by training samples. This threshold is rescaled back to 0.5 in Figure 10. Two peaks at the midnight 29 and 30 October can be well captured by these predicted probabilities. The time shift between the peak of real Dst and the peak of the predicted probability is no more than 4 hr. This implies that those strong storm can predicted very well by the proposed model.

Figure 10 also indicates that the proposed model can forecast well those storms caused by CME because they stand out very clearly in SoHO images. However, a good portion of solar CMEs are non-Earth oriented. The ability of this model to identify the geo-effectiveness (or lack thereof) of non-Earth-directed CMEs is also assessed. Twenty of them occurring in the period 2000–2003 are selected for validation. The TSS and MCC of the prediction based on the developed model during those non-Earth-directed CME periods are 0.94 and 0.46 respectively. The TP, FP, TN and FN values are 14, 50, 709, and 0. An example is shown in Figure 11. A strong CME occurred around 30 July 2002, but the corresponding Dst did not reach $-100$ nT. The predicted Dst probability increases but not as significantly as for the Earth-directed CMEs. This suggests that the proposed model may be able to distinguish non-Earth-directed CMEs, and assign lower Dst probabilities to them. Similar plots of probabilities for all the other storm events used in this study are included as supplementary information.

| Table 7
| Accuracy of the Developed Model With Different Time Resolution |
|-----------------|-------|-----|-----|-----|-----|
| Time resolution | TSS   | MCC | TP  | FP  | TN  | FN  |
| 2               | 0.62  | 0.37| 57  | 209 | 1738| 21  |
| 6               | 0.42  | 0.25| 9   | 47  | 555 | 9   |

TSS, True skill statistic; MCC, Matthews Correlation Coefficient; TP, true positive; FP, false positive; TN, true negative; FN, false negative numbers.

| Table 8
| Weighted Metrics of the Developed Model With Different Time Resolution |
|-----------------|-------|-----|-----|-----|-----|
| Time res        | wTSS | wMCC| wTP | wFP | wTN | wFN |
| 2               | 0.68 | 0.47| 67  | 133 | 1407| 20  |
| 6               | 0.59 | 0.42| 41  | 82  | 510 | 15  |
5. Summary and Outlook

We have developed a LS-based class-balanced ensemble CNN model that estimates the probability of $Dst$ exceeding a given threshold 1 day ahead based on SoHO images. Fifty-one selected storm events were chosen during a long-span historical data set (~16 years), between 01 May 1996 and 20 April 2011. The proposed model can

Figure 10. Probabilities generated from the developed model during 25 October 2003 and 05 November 2003, together with the corresponding $Dst$ and the thresholds for both probabilities and $Dst$ during this storm cases. x axis is Date. Left and right Y axes are probability (shown by green line) and $Dst$ (shown by blue line) respectively. Cross line is when the probability equals to 0.5, and dash line is when $Dst$ equals to $-100$ nT. Corresponding images used for training and prediction are plotted in the bottom panel. From top to bottom, they are for Michelson Doppler Imager (MDI), Extreme ultraviolet Imaging Telescope-195 and Large Angle and Spectrometric Coronagraph C2.
predict the probability that $D_{st} < -100$ nT 24 hr ahead with a TSS of 0.62 and MCC of 0.37. The weighted TSS and MCC from Guastavino et al. (2021) are 0.68 and 0.47.

One of the crucial points of this work is that it combines a LS ensemble method with a CNN algorithm for a probability prediction. A customized class-balanced mean square error is developed as the cost function of the proposed CNN model. After the CNN model is developed, a LS method is developed to estimate the weights of the predictions from the proposed multi-target CNNs. Eventually a final probability value can be calculated by

Figure 11. Similar to Figure 10. An example during an non-Earth-direct CME during 31 July 2002 and 02 August 2002.
the optimized weights and CNN predictions. Binary classification of each event is then determined by a threshold re-calibrated on the other 50 storms used for training.

We have shown that this proposed model provides good skills for predicting $Dst$ 1-day-ahead during strong storm periods. The proposed model can also forecast $Dst$ probability even within a non-Earth-direct CME period. This model will extend the prediction lead time of most of the current $Dst$ empirical prediction models. The performance metrics that we have analyzed are the confusion matrix, TSS, MCC, and corresponding weighted scores from Guastavino et al. (2021). Finally, we have discussed a strong storm case from 25 October 2003 to 05 November 2003. $Dst$ peaks can be well captured by the developed model.

A possible weakness of this model is that the weights are static, although the time lags from different events should not be the same (Chandorkar et al., 2019). As a next step, we plan to take into account dynamic weights, for example, by applying an online/dynamic ensemble method, such as Monteleoni et al. (2011). Moreover, storms that result from CMEs or high speed streamers are based on different physical mechanisms. The former mostly occur during high solar activity period, while the latter are seen more often during low solar activity period. Including a solar activity index, such as $F_{10.7}$ into consideration to first classify those samples in order to train the model more precisely, will be also experimented in the future.

Data Availability Statement

We thank OMNIWeb for providing the $Dst$ data (https://omniweb.gsfc.nasa.gov/), the NASA Solar Data Analysis Center’s (SDAC) Virtual Solar Observatory (VSO) (https://sdac.virtualsolar.org/cgi/search) and Stanford University’s Joint Science Operation Center (JSOC) http://jsoc.stanford.edu/MDI/MDI_Magneto.png for SoHO images. All the results and codes have been made available as a Zenodo repository in https://doi.org/10.5281/zenodo.6385325. Future updates can be found on https://github.com/HuanWinter/Dst_SoHO and https://ml-space-weather.github.io/projects.html (under construction).

Acknowledgments

This project has been developed in the framework of the European project AIDA. The AIDA project has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 776262. EC is partially funded by the AIDA. The AIDA project has received this project has been developed in the University’s Joint Science Operation Center (JSOC) and Stanford University’s Joint Science Operation Center’s (SDAC) Virtual Solar Observatory (VSO) (under construction).

References

Awad, M., & Khanna, R. (2015). Support vector regression. In Efficient learning machines (pp. 67–80). Springer. https://doi.org/10.1007/978-1-4302-5989-9_4

Bala, R., & Reiff, P. (2012). Improvements in short-term forecasting of geomagnetic activity. Space Weather, 10(6). https://doi.org/10.1029/2012sw000779

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics, 16(5), 412–424. https://doi.org/10.1093/bioinformatics/16.5.412

Basso, C. D., & Ramos, A. A. (2018). Enhancing SDO/HMI images using deep learning. Astronomy & Astrophysics, 614, A5. https://doi.org/10.1051/0004-6361/201731344

Borovsky, J. E., & Shprits, Y. Y. (2017). Is the Dst index sufficient to define all geospace storms? Journal of Geophysical Research: Space Physics, 122(11), 1–543. https://doi.org/10.1002/2017ja024679

Bowman, B., Tobiska, W. K., Marcos, F., Huang, C., Lin, C., & Burke, W. (2008). A new empirical thermospheric density model JB2008 using new solar and geomagnetic indices. In AIAA/AAS astrodynamics specialist conference and exhibit (p. 6438).

Brueckner, G. E., Howard, R. A., Koomen, M. J., Korendyke, C. M., Michels, D. J., Moses, J. D., et al. (1995). The Large Angle spectroscopic Coronagraph (LASC). Solar Physics, 162(1–2), 357–402. https://doi.org/10.1007/BF00733434

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. Space Weather, 17(8), 1166–1207. https://doi.org/10.1029/2018sw002061

Camporeale, E., Cash, M. D., Singer, H. J., Balch, C. C., Huang, Z., & Toth, G. (2020). A gray-box model for a probabilistic estimate of regional ground magnetic perturbations: Enhancing the NOAA operational geospace model with machine learning. Journal of Geophysical Research: Space Physics, 125(11), e2019JA027884. https://doi.org/10.1029/2019ja027884

Chandorkar, M., & Camporeale, E. (2018). Probabilistic forecasting of geomagnetic indices using Gaussian process models. In Machine learning techniques for space weather (pp. 237–258). Elsevier. https://doi.org/10.1016/b978-0-12-817788-0.00009-3

Chandorkar, M., Camporeale, E., & Wing, S. (2017). Probabilistic forecast of the disturbance storm time index: An autoregressive Gaussian process approach. Space Weather, 15(8), 1004–1019. https://doi.org/10.1002/2017sw001627

Chandorkar, M., Furtlechner, C., Poduval, B., Camporeale, E., & Sebag, M. (2019). Dynamic time lag regression: Predicting what & when. In International conference on learning representations.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., & Belongie, S. (2019). Class-balaneced loss based on effective number of samples. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9268–9277).

Delaboudinière, J. P., Artzner, G. E., Brunaud, J., Gabriel, A. H., Hochedez, J. F., Millier, F., et al. (1995). EIT: Extreme-ultraviolet imaging telescope for the SOHO mission. Solar Physics, 162(1–2), 291–312. https://doi.org/10.1007/BF00733432

Domingo, V., Fleck, B., & Poland, A. I. (1995). The SOHO mission: An overview. Solar Physics, 162(1–2), 1–37. https://doi.org/10.1007/BF00733425

Dos Santos, L. F., Bose, S., Salvatelli, V., Neuberg, B., Cheng, M. C., Janvier, M., et al. (2021). Multichannel autocalibration for the atmospheric imaging assembly using machine learning. Astronomy & Astrophysics, 648, A53. https://doi.org/10.1051/0004-6361/202040051
Gruet, M. A., Chandokkar, M., Sicard, A., & Camporeale, E. (2018). Multiple-hour-ahead forecast of theDstindex using a combination of long short-term memory neural network and Gaussian process. *Space Weather, 16*(11), 1882–1896. https://doi.org/10.1029/2018sw001898

Guastavino, S., Piana, M., & Benvenuto, F. (2021). *Bad and good errors: Value-weighted skill score in deep ensemble learning*. arXiv preprint arXiv:2103.02881.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., et al. (2018). Recent advances in convolutional neural networks. *Pattern Recognition, 77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

Hu, A., Sisti, M., Finelli, F., Califano, F., Dargent, J., Faganello, M., et al. (2020). Identifying magnetic reconnection in 2D Hybrid Vlasov Maxwell simulations with convolutional neural networks. *The Astrophysical Journal, 900*(1), 86. https://doi.org/10.3847/1538-4357/aba527

Illarionov, E. A., & Tlatov, A. G. (2018). Segmentation of coronal holes in solar disc images with a convolutional neural network. *Monthly Notices of the Royal Astronomical Society, 481*(4), 5014–5021. https://doi.org/10.1093/mnras/sty2628

Laperre, B., Amaya, J., & Lapenta, G. (2020). *Dynamic time warping as a new evaluation for Dst forecast with machine learning* (Vol. 7). arXiv preprint arXiv:2006.04667. https://doi.org/10.3389/fspas.2020.00039

Lazzòs, J., Vega, P., Rojas, P., & Sallate, I. (2017). Forecasting the Dst index using a swarm-optimized neural network. *Space Weather, 15*(8), 1068–1089. https://doi.org/10.1002/2017sw001608

Licata, R. J., Tobiska, W. K., & Mehta, P. M. (2020). Benchmarking forecasting models for space weather drivers. *Space Weather, 18*(10), e2020SW002496. https://doi.org/10.1029/2020sw002496

Liemohn, M. W., Koszyra, J., Thomsen, M., Roeder, J., Lu, G., Borovsky, J., & Cayton, T. (2001). *Dominant role of the asymmetric ring current in producing the stormtime Dst*. *Journal of Geophysical Research, 106*(A6), 10883–10904. https://doi.org/10.1029/2000ja000326

Li, X., Zheng, Y., Wang, X., & Wang, L. (2020). Predicting solar flares using a novel deep convolutional neural network. *The Astrophysical Journal, 891*(1), 10. https://doi.org/10.3847/1538-4357/ab6d04

Lundstedt, H., Gleisner, H., & Wintoft, P. (2002). *Operational forecasts of the geomagnetic Dst index*. *Geophysical Research Letters, 29*(24), 34–41. https://doi.org/10.1029/2002glb106151

Monteleoni, C., Schmidt, G. A., Saroja, S., & Asplund, E. (2011). Tracking climate models. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 4*(4), 372–392. https://doi.org/10.1002/sam.10126

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., & Shin, G. (2018). Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms. *The Astrophysical Journal, 869*(2), 91. https://doi.org/10.3847/1538-4357/aa6b04

Polikar, R. (2012). Ensemble learning. In *Ensemble machine learning* (pp. 1–34). Springer. https://doi.org/10.1007/978-1-4419-9326-7_1

Rostoker, G. (1972). *Geomagnetic indices*. *Reviews of Geophysics, 10*(4), 935–950. https://doi.org/10.1029/RG010i004p00935

Ruwali, A., Kumar, A. S., Prakash, K. B., Sivavaraprasad, G., & Ratnam, D. V. (2020). Implementation of hybrid deep learning model (LSTM-CNN) for ionospheric TEC forecasting using GPS data. *IEEE Geoscience and Remote Sensing Letters, 18*(6), 1004–1008. https://doi.org/10.1109/lgrs.2020.2992633

Saiz, E., Cid, C., & Cerrato, Y. (2008). Forecasting intense geomagnetic activity using interplanetary magnetic field data. *Annales geophysicae, 26*(12), 3989–3998. https://doi.org/10.5194/angeo-26-3989-2008

Scherrer, P. H., Bogart, R. S., Bush, R. I., Hoeksema, J. T., Kosovichev, A. G., Schou, J., et al. (1995). The solar oscillations investigation—Michelson Doppler imager. *Solar Physics, 162*(1–2), 129–188. https://doi.org/10.1007/BF00733429

Shneider, C., Hu, A., Tiwari, A. K., Bobra, M. G., Battams, K., Teunissen, J., & Camporeale, E. (2021). *A machine-learning-ready Dataset prepared from the solar and heliospheric observatory mission*. arXiv preprint arXiv:2108.06394.

Siciliano, F., Consolini, G., Tozzi, R., Gentili, M., Giannattasio, F., & De Michelis, P. (2021). Forecasting sym-h index: A comparison between long short-term memory and convolutional neural networks. *Space Weather, 19*(2), e2020SW002589. https://doi.org/10.1029/2020sw002589

Sugiura, M., & Kamei, T. (1991). *Equatorial Dst index 1957–1986*, IAGA bull., 1991. In A.Berthelier, & M.Menville (Eds.), (Int. Serv. Geomagn.

Tobiska, W., Knipp, D., Burke, W., Bouwer, D., Bailey, J., Odstrcil, D., et al. (2013). The anemomilos prediction methodology for Dst. *Space Weather, 11*(9), 490–508. https://doi.org/10.1002/swe.20094

Upendra, V., Cheng, M. C., Hansagose, S., & Krishnamurthi, G. (2020). Solar wind prediction using deep learning. *Space Weather, 18*(9), e2020SW002478. https://doi.org/10.1029/2020sw002478

Xu, S., Huang, S., Yuan, Z., Deng, X., & Jiang, K. (2020). Prediction of the Dst index with bagging ensemble-learning algorithm. *The Astrophysical Journal: Supplement Series, 248*(1), 14. https://doi.org/10.3847/1538-4365/ab801b

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x