Deep-learning Reconstruction of Sunspot Vector Magnetic Fields for Forecasting Solar Storms

Dattaraj B. Dhuri1,2, Shamik Bhattacharjee1, Shravan M. Hanasoge1,2, and Sashi Kiran Mahapatra1
1 Department of Astronomy and Astrophysics, Tata Institute of Fundamental Research, Mumbai, 400005 India; dbd7602@nyu.edu
2 Center for Space Science, New York University Abu Dhabi, PO Box 129188, Abu Dhabi, UAE

Abstract

Solar magnetic activity produces extreme solar flares and coronal mass ejections, which pose grave threats to electronic infrastructure and can significantly disrupt economic activity. It is therefore important to appreciate the triggers of explosive solar activity and develop reliable space weather forecasting. Photospheric vector magnetic field data capture sunspot magnetic field complexity and can therefore improve the quality of space weather prediction. However, state-of-the-art vector field observations are consistently only available from Solar Dynamics Observatory/Helioseismic and Magnetic Imager (HMI) since 2010, with most other current and past missions and observational facilities, such as Global Oscillations Network Group (GONG), only recording line-of-sight (LOS) fields. Here, using an inception-based convolutional neural network (CNN), we reconstruct HMI sunspot vector field features from LOS magnetograms of HMI and GONG with high fidelity (~90% correlation) and sustained flare forecasting accuracy. We rebuild vector field features during the 2003 Halloween storms, for which only LOS field observations are available, and the CNN-estimated electric current helicity accurately captures the observed rotation of the associated sunspot prior to the extreme flares, showing a striking increase. Our study thus paves the way for reconstructing three solar cycles worth of vector field data from past LOS measurements, which are of great utility in improving space weather forecasting models and gaining new insights about solar activity.

Unified Astronomy Thesaurus concepts: Solar activity (1475); Solar active region magnetic fields (1975)

Supporting material: figure set

1. Introduction

Sunspot magnetic fields are generated within the solar interior, become buoyant through the solar convection zone, and emerge at the photosphere and the corona as large-scale structures of sunspots and active regions (ARs) in the form of giant loops (Cheung & Isobe 2014). Coronal loops are dynamic, driven by emerging magnetic flux, electric current, and turbulent flows. Free magnetic energy stored in these loops is occasionally released via magnetic reconnection in the form of explosions such as flares and coronal mass ejections (CMEs; Shibata & Magara 2011; Su et al. 2013). Radiation and charged particles emitted in these explosions can lead to severe space weather, disrupting our life on Earth significantly (Pulkkinen et al. 2005; Eastwood et al. 2017; Boteler 2019). In the past, the geomagnetic storm of 1989, resulting from an X15-class flare and subsequent CME, tripped circuit breakers in the Hydro-Quebec power grid, causing a widespread blackout in Quebec (Boteler 2019). The Halloween storm of 2003 produced extreme flares, causing transformer malfunction and blackouts in Sweden and damaging multiple science-mission satellites (Pulkkinen et al. 2005). In today’s society, a high-magnitude solar storm can potentially lead to trillions of US dollars worth of economic losses, with up to a decade of recovery time (Eastwood et al. 2017). Improving our understanding of AR magnetic fields is therefore important for identifying triggers of these explosions and achieving reliable space weather forecasting.

Coronal and photospheric AR magnetic fields are non-potential, comprising twisted flux tubes as revealed by high-resolution, high-cadence observations of the Solar Dynamics Observatory (SDO; Pesnell et al. 2012) since 2010. Large ARs and their complex dynamics, e.g., twisting and rotation, are known to be associated with solar explosive activity (Toriumi & Wang 2019). The SDO/Helioseismic and Magnetic Imager (HMI) photospheric vector magnetic field observations facilitate the calculation of AR features (Leka & Barnes 2007), such as total unsigned magnetic flux, free energy density, electric current helicity, and Lorentz forces, characterizing the AR magnetic field dynamics. These features are publicly available as the HMI data product Space-weather HMI Active Region Patch (SHARP; Bobra et al. 2014). The SHARP features are extensively used for statistical studies of preflare magnetic field evolution and energy buildup (Dhuri et al. 2019) and improving space weather forecasting using machine learning (ML; Bobra & Couvidat 2015; Bobra & Ilonidis 2016; Chen et al. 2019). HMI observations are limited to only one full solar cycle (cycle 24); therefore, statistical space weather forecasting models based on SHARP are restricted.

Various difficulties are associated with the measurement of the transverse component of the photospheric magnetic field (Stenflo 2013); therefore, ground- and space-based instruments monitoring the Sun since the 1970s provide observations of only the longitudinal, i.e., line-of-sight (LOS), component. Continuous full-disk LOS field observations are available through the ground-based NASA/National Solar Observatory (NSO) Kitt Peak Telescope (1974–present) (Livingston et al. 1976), space-based Michelson and Doppler Imager (MDI, 1996–2011; Scherrer et al. 1995), and ground-based Global Oscillations Network Group (GONG, 1995–present). These
LOS field measurements, although not sufficient for quantifying sunspot complexity to nonpotential energy and helicity, have been useful for providing a qualitative assessment of AR morphology via sunspot classification schemes, such as the McIntosh classification (McIntosh 1990) and Mount Wilson classification, which form the basis of operational space weather forecasts (Crown 2012).

Improving on these tentative AR classifications and formally devising a method to quantify vector field properties from LOS fields is of great utility because (i) it allows for “improving” past data sets of LOS observations and understanding how vector field features have evolved over multiple solar cycles, (ii) a reliable estimation of vector field features over the past few decades can be used to build more robust statistical models for space weather forecasting, and (iii) for future missions acquiring only LOS data, vector field features and even full vector field construction can be part of an on-ground data-processing pipeline. ML methods such as convolutional neural networks (CNNs) developed through the past decade have proven to be hugely successful in identifying patterns and correlations in large, high-dimensional data sets and particularly images (LeCun et al. 2015; Goodfellow et al. 2016). Here we explore dependencies between LOS magnetograms and the corresponding full vector field of ARs through a CNN model developed to estimate vector field features of SHARP using the LOS magnetograms measurements from space-based SDO/HMI and ground-based GONG.

2. Data

We use photospheric LOS magnetogram data provided by HMI and GONG. GONG provides only LOS magnetograms. HMI-derived SHARP (the hmi.sharp_cea_720s data series; Bobra et al. 2014) includes vector and LOS magnetograms of AR patches that are automatically detected and tracked as they rotate across the visible solar disk (Bobra et al. 2014). HMI magnetograms are available at a plate scale of 0.5°, i.e., ~380 km at the disk center. GONG magnetograms are available at a plate scale of 2.5°. The magnetograms available in the hmi.sharp_cea_720s series are on a cylindrical equal-area (CEA) grid, thus eliminating the projection effects. We similarly remap the GONG AR magnetograms to a CEA grid. We train a CNN to obtain SHARP features directly from LOS magnetograms of HMI and GONG. We only consider top SHARP features that produce maximum flare forecasting accuracy for an ML model (Bobra & Couvidat 2015). These are listed in Table 2.

HMI measurements are sensitive to the observation conditions and the relative velocity between SDO and the Sun (Hoeksema et al. 2014). Observation conditions are indicated by the QUALITY flag, and we consider measurements for which the Stokes vectors are reliable (QUALITY ≤ 10,000 in hexadecimal) and when the relative velocity between SDO and the Sun is <3500 m s⁻¹ (Bobra & Ilonidis 2016). Data closer to the limb are noisier because of the higher relative velocities and projection effects. Therefore, we limit observations to within ±45° of the central meridian. Further, we only include ARs from the SHARP data series that grow to a maximum area of >25 Mm². This eliminates a significant number of small ARs that do not produce major (M- or X-class) flares. The SHARP feature calculation using HMI vector field observations considers those pixels in the AR magnetograms for which the 180° ambiguity resolution is reliable (Bobra et al. 2014).

| # HMI ARs | 848 | 194 |
| # HMI samples | 124,633 | 26,820 |
| # GONG ARs | 848 | 145 |
| # GONG samples | 114,443 | 13,454 |

Observations between 2010 May and 2018 August are used to train the CNN—approximately 80% of the data are used to train and validate the CNN, while the remainder are the unseen or test data. We chronologically split the ARs into training and validation and test partitions: ARs in the period 2010 May–2015 September are for training and validation, and 2015 October–2018 August is for the test. Samples are drawn every 6 hr from the time series of each AR. All samples from a given AR are exclusively part of the training set, the validation set, or the test set, to avoid biases arising from temporal coherence of observations from an AR (Ahmazadeh et al. 2021). The numbers of ARs and magnetogram observations used for training, validation, and test are listed in Table 1. Since solar activity depends on the phase of the cycle, the chronological splitting may introduce a bias for training the CNN. Indeed, the ratio of flaring to nonflaring ARs in the test data is approximately half its value in the training and validation data set (Bhattacharjee et al. 2020). However, chronological splitting is appropriate for operational space weather forecasting tools.

3. Methods

CNNs are neural networks with convolution filters (kernels) to scan over the input data, typically 2D data of images, and detect spatial patterns for tasks such as classification and identification (LeCun et al. 2015; Goodfellow et al. 2016). The convolution filters are K × K neurons that slide over the images and detect different patterns. Convolution filters have free parameters—each neuron has weight w, and each convolution filter has bias b. Neurons process pixels of the inputs (or the outputs from previous layers) xᵢ by performing the operation f(∑ᵢ wᵢxᵢ + b), where f is the activation function (Hastie et al. 2001). CNNs also have pooling layers that are used to reduce the input size as it progresses to deeper levels of the CNN. A max- or average-pooling filter picks out the maximum or average value from the given N × N feature map. Pooling layers typically follow a convolutional layer in a CNN to reduce the dimensionality.

We use a CNN architecture with inception modules similar to inception V1 modules from GoogleNet (Szegedy et al. 2015). Typically, in a convolutional layer, we use filters of fixed size that work best for the particular problem. However, inception modules are designed to detect patterns over a variety of length scales that may be present in the input. They involve convolution filters of different sizes in a single layer. The outputs from all the convolutional layers in an inception module are concatenated and supplied as an input to the following layer. The inception module used here comprises
three convolution filters of sizes $3 \times 3$, $5 \times 5$, and $7 \times 7$ and one $3 \times 3$ max-pooling filter.

The CNN architecture is shown in Figure 1. The CNN takes in two inputs: (i) LOS magnetograms of AR patches, and (ii) AR center latitude ($\lambda_c$) and longitude ($\phi_c$). There are no fully connected layers that directly process the magnetogram input, and therefore the CNN can process magnetogram patches of variable sizes. (b) Inception module used in the CNN.

![Figure 1. The CNN architecture. (a) CNN architecture used for obtaining vector field features from LOS magnetograms. The architecture incorporates inception modules similar to GoogleNet (Szegedy et al. 2015). The CNN takes in two inputs: (i) LOS magnetograms, and (ii) AR center latitude ($\lambda_c$) and longitude ($\phi_c$). There are no fully connected layers that directly process the magnetogram input, and therefore the CNN can process magnetogram patches of variable sizes. (b) Inception module used in the CNN.](image)

Figure 2. Schematic of ML models. Left: CNN models process LOS magnetograms as input and produce vector field features of SHARP as output. SHARP features group together in four groups based on their mutual correlations as shown. We develop four different CNN models to estimate SHARP from four different groups. All CNN models have identical architecture described in Figure 1 except the final output layer, where the number of output neurons is equal to the number of SHARP features to be estimated from the respective group. Right: baseline models using LR for estimation of two groups of SHARP features that depend on electric current and free energy, respectively, using extensive SHARP features and Schrijver’s $R$ value (Schrijver 2007) as an input.

![Figure 2. Schematic of ML models. Left: CNN models process LOS magnetograms as input and produce vector field features of SHARP as output. SHARP features group together in four groups based on their mutual correlations as shown. We develop four different CNN models to estimate SHARP from four different groups. All CNN models have identical architecture described in Figure 1 except the final output layer, where the number of output neurons is equal to the number of SHARP features to be estimated from the respective group. Right: baseline models using LR for estimation of two groups of SHARP features that depend on electric current and free energy, respectively, using extensive SHARP features and Schrijver’s $R$ value (Schrijver 2007) as an input.](image)
the convolutional layers, which explicitly treats the positive and negative pixel values from LOS magnetograms symmetrically. In addition, the fully connected layer of neurons has a \( \text{tanh} \) activation to explicitly treat positive and negative values of latitude and longitude, which are normalized between \([-1, 1]\), symmetrically. The final output layer of neurons has \( \text{sigmoid} \) activation (Han & Moraga 1995; Hastie et al. 2001) to yield the normalized value of the estimated SHARP features between \(0\) and \(1\).

The absence of fully connected layers in the network that processes the LOS magnetogram input implies that the CNN architecture can analyze LOS magnetograms of arbitrary sizes. Since AR patches are of varied dimensions, magnetograms in the training, validation, and test data are also correspondingly differently sized. As such, our CNN does not require preprocessing to convert magnetograms to a fixed size, and thus it is free from biases that may arise as a result of resizing (Bhattacharjee et al. 2020).

We use 10 times repeated-holdout validation for training the CNN (Hastie et al. 2001). We randomly split ARs in the training and validation sets into three parts and use data from two parts for training and the remaining part for validation. This process is repeated nine times while ensuring that the data from an AR are part of either the training or the validation and not both. The output from the CNN is compared to the original SHARP feature values. The \( \text{sigmoid} \) output layer of the CNN lies in a continuous range between \(0\) and \(1\). The original SHARP features are normalized by dividing by their respective maximum values. We partition the normalized features (over the range from \(0\) to \(1\)) in the training set into 10 bins of equal width (0.1) and oversample the data in each bin to match the number of samples in the maximally populated bin. The input magnetograms are standardized, i.e., a mean is subtracted and the resultant magnetogram is divided by a standard deviation of the magnetic field values. The mean and standard deviation used for standardization are calculated over all pixels of all magnetograms in the training and validation data of the respective instrument. The CNN output is compared to the original SHARP values, and the loss function—defined to be the mean squared error—is computed. We train the CNN to minimize the mean squared error(802,839),(944,847) to achieve a learning rate of \(0.00007\). The CNN is developed using the Python library \textit{keras}.

### 4. Results

#### 4.1. Estimation of AR Vector Magnetic Field Features Using CNN

The SHARP features considered (listed in Figure 2 and Table 2) are correlated among each other and are divided into groups based on their Pearson correlations (Bobra et al. 2014): (i) features that depend on the area of ARs, i.e., extensive features that include AR area, total unsigned flux, total unsigned vertical current, total unsigned current helicity, total free energy density, and total Lorentz force; (ii) features that depend on the electric current in ARs, i.e., absolute net current helicity and sum of net current per polarity; (iii) features that depend only on the nonpotential energy in ARs, i.e., mean free energy density, and (iv) \( \text{R}_\text{value} \), i.e., the sum of flux near the polarity inversion line (Schrijver 2007).

The SHARP features are strongly correlated with AR total unsigned flux, which depends only on the radial component of the magnetic field. The radial component is traditionally estimated from AR LOS magnetic field using a potential field approximation (Leka et al. 2017). Using a CNN, we directly estimate these extensive features without first requiring to estimate the radial magnetic field. The \( \text{R}_\text{value} \) depends only on LOS magnetic field and can be directly calculated using GONG LOS magnetograms. However, to match HMI SHARP \( \text{R}_\text{value} \), GONG LOS magnetic fields require a cross-calibration. CNN models are expected to implicitly learn the cross-instrument calibration during training (Munoz-Jaramillo et al. 2022), and estimated SHARP values are also expected to be automatically cross-calibrated.

Unlike the extensive features, an accurate estimation of SHARP depending on electric current and mean free energy

---

**Table 2**

Pearson Correlations between the CNN-estimated Vector Field Features of SHARP and Their True Values

| SHARP Features                        | HMI          | GONG         | HMI          | GONG         |
|---------------------------------------|--------------|--------------|--------------|--------------|
| Total unsigned flux                   | 95.14 ± 00.62| 90.87 ± 01.96| 89.73 ± 02.70| 87.42 ± 01.39|
| Area                                  | 95.87 ± 00.49| 95.06 ± 00.84| 92.00 ± 01.70| 92.88 ± 00.88|
| Total unsigned vertical current       | 94.78 ± 00.71| 91.80 ± 01.69| 88.86 ± 02.57| 89.00 ± 01.76|
| Total unsigned current helicity       | 95.74 ± 00.50| 91.65 ± 01.76| 88.33 ± 02.65| 83.31 ± 02.28|
| Total free energy density             | 96.19 ± 00.80| 92.60 ± 01.60| 90.17 ± 02.37| 91.22 ± 01.25|
| Total Lorentz force                   | 96.64 ± 00.47| 94.94 ± 00.98| 90.63 ± 02.46| 92.71 ± 00.87|
| Absolute net current helicity         | 90.37 ± 03.28| 63.76 ± 03.65| 57.83 ± 08.84| 57.60 ± 06.97|
| Sum of net current per polarity       | 89.51 ± 02.53| 64.58 ± 03.08| 61.93 ± 07.63| 59.09 ± 06.96|
| Mean free energy density              | 95.10 ± 01.00| 89.92 ± 00.79| 92.13 ± 01.80| 91.73 ± 00.54|
| Area with shear >45°                  | 95.02 ± 00.81| 90.00 ± 01.19| 90.59 ± 01.57| 90.48 ± 00.46|
| Flux near polarity inversion line     | 90.54 ± 00.56| 76.28 ± 01.83| 77.11 ± 00.79| 70.43 ± 00.79|

**Note.** SHARP features are calculated from HMI vector field observations (Bobra et al. 2014). The \( p \)-values for all correlations are \( \sim 0.0 \). SHARP features are mutually correlated (Dhuri et al. 2019) and accordingly arranged in the four groups as features depending on (i) AR area, (ii) electric current, (iii) mean free energy density, and (iv) \( \text{R}_\text{value} \), i.e., the sum of flux near the polarity inversion line (Schrijver 2007).
Two baseline LR models are developed as a baseline. We develop two separate LR models, one that takes extensive SHARP features and Schrijver’s $R_{value}$ as input and produce SHARP features that depend on electric current and mean nonpotential energy as output. The $p$-values for all correlations are $\sim 0.0$.

requires explicit knowledge of the full vector magnetic fields. Such features are important for understanding triggers of solar storms and are typically estimated assuming magnetic field models, e.g., linear and nonlinear force-free models (Régnier & Priest 2007). Here we provide a purely data-driven estimation of these features using a CNN. In order to assess the performance of the CNN, we use linear regression (LR) models as a baseline. We develop two separate LR models, one each for features that depend on electric current and free energy, respectively. As input, the LR models have extensive features and $R_{value}$. The first LR model produces absolute net current helicity and the sum of net current per polarity as the output, while the second one produces mean free energy density and area with shear $>45^\circ$ as the output. Figure 2 shows a schematic of the CNN models and the baseline models.

We use Pearson and Spearman correlations for measuring the performance of the CNN and baseline models. Pearson correlation measures a linear correlation between the true and estimated values of the vector magnetic field features. Spearman correlation is a rank correlation that captures the monotonic relationship between the true and estimated values in addition to the linear relationship measured by the Pearson correlation. Pearson and Spearman correlations for the CNN-estimated vector magnetic field features are listed in Tables 2 and 3, respectively. For the baseline models, these correlations are listed in Table 4.

From Table 2, the Pearson correlations of CNN-estimated vector field features are higher for HMI than GONG and thus appear to be dependent on the spatial resolution of LOS magnetograms. For HMI, the CNN-estimated extensive features yield a Pearson correlation of $\sim 95\%$ for the validation data and $\sim 90\%$ for the test data. For GONG data, the corresponding correlation is $\sim 90\%$. The Pearson correlations of CNN-estimated values of extensive features are not a perfect $\sim 100\%$, since the SHARP calculation does not consider all pixels, but rather only taking into account those for which disambiguation of the azimuthal component of the magnetic field is reliable (Bobra et al. 2014). From Table 3, the Spearman correlations for the extensive features are only slightly lower than the corresponding Pearson correlations, implying that the ranking of the estimated features is generally consistent with the true ranking.

The Pearson and Spearman correlation values for features that depend on the nonpotential energy are significantly high ($>90\%$) across the validation and test data sets. These correlation values are also $\sim 10\%$–$20\%$ higher compared to the LR baseline model (Table 4). These features, namely, mean free energy density and area with shear $>45^\circ$, explicitly depend on the full vector magnetic field.

For features that depend on electric current—i.e., absolute net current helicity and sum of net current per polarity—the CNN does not perform better than the baseline model. While the Pearson correlations for the validation are $20\%$ higher compared to the baseline of $70\%$, Spearman correlations are approximately equal (up to the error bars) at $60\%$. In addition, the CNN fails to generalize to the test data, with low Pearson and Spearman correlation scores of $60\%$ each.

Note. The $p$-values for all correlations are $\sim 0.0$.

| SHARP Features                              | 10 times Repeated-holdout Validation | Test     |
|---------------------------------------------|--------------------------------------|----------|
| Total unsigned flux                         | 86.92 ± 0.13                        | 81.19 ± 0.74 |
| Area                                        | 89.57 ± 0.10                        | 87.10 ± 0.33 |
| Total unsigned vertical current             | 86.82 ± 0.41                        | 81.22 ± 0.10 |
| Total unsigned current helicity              | 87.37 ± 0.41                        | 85.53 ± 0.24 |
| Total free energy density                   | 84.23 ± 0.14                        | 81.32 ± 0.33 |
| Total Lorentz force                         | 90.63 ± 0.04                        | 86.96 ± 0.51 |
| Absolute net current helicity               | 59.61 ± 0.26                        | 57.70 ± 0.69 |
| Sum of net current per polarity             | 60.02 ± 0.95                        | 57.51 ± 0.25 |
| Mean free energy density                    | 92.02 ± 0.97                        | 93.02 ± 0.42 |
| Area with shear $>45^\circ$                 | 93.19 ± 0.98                        | 91.63 ± 0.65 |
| Flux near polarity inversion line           | 91.69 ± 0.47                        | 76.63 ± 0.60 |

Note. Two baseline LR models are developed (Figure 2) that take extensive SHARP features and Schrijver’s $R_{value}$ as input and produce SHARP features that depend on electric current and mean nonpotential energy as output. The $p$-values for all correlations are $\sim 0.0$.

| SHARP Features                              | Pearson | Spearman | Pearson | Spearman |
|---------------------------------------------|---------|----------|---------|----------|
| Total unsigned flux                         | 86.27 ± 0.75 | 81.19 ± 0.74 | 77.11 ± 0.20 |
| Area                                        | 92.11 ± 0.75 | 87.10 ± 0.33 | 86.91 ± 0.21 |
| Total unsigned vertical current             | 87.34 ± 0.38 | 81.22 ± 0.10 | 79.07 ± 0.52 |
| Total unsigned current helicity              | 87.49 ± 0.48 | 82.61 ± 0.24 | 79.40 ± 0.26 |
| Total free energy density                   | 85.53 ± 0.17 | 81.32 ± 0.27 | 80.11 ± 0.24 |
| Total Lorentz force                         | 92.78 ± 0.09 | 86.96 ± 0.51 | 86.49 ± 0.16 |
| Absolute net current helicity               | 59.35 ± 0.36 | 57.70 ± 0.69 | 42.27 ± 0.73 |
| Sum of net current per polarity             | 65.75 ± 0.95 | 57.51 ± 0.25 | 47.85 ± 0.79 |
| Mean free energy density                    | 91.59 ± 0.92 | 93.02 ± 0.42 | 92.58 ± 0.35 |
| Area with shear $>45^\circ$                 | 90.08 ± 0.54 | 91.63 ± 0.65 | 89.96 ± 0.37 |
| Flux near polarity inversion line           | 76.63 ± 0.60 | 83.00 ± 0.85 | 75.32 ± 0.92 |

Note. The $p$-values for all correlations are $\sim 0.0$.
Figure 3 shows scatter plot visualizations of the correlation between true and CNN-estimated SHARP features for HMI and GONG from the 10 validation sets. For HMI, the true and CNN-derived values mostly match relatively closely, except at only very small values (<200 \( G^2 - m^{-1} \)) of absolute net current helicity, where the CNN estimates are significantly larger. For the HMI test data and GONG data, the CNN estimation of absolute net current helicity for large values (>1000 \( G^2 - m^{-1} \)) is consistently on the lower side (~500 \( G^2 - m^{-1} \)). Figure 4 explicitly shows mean absolute errors in the CNN estimation as a function of the true values for HMI and GONG. Mean absolute errors in CNN-estimated values from GONG magnetograms show higher dependence on true values compared to HMI and increase significantly with increasing true values of the respective features, particularly for the validation data. For total unsigned flux, mean absolute errors of CNN-estimated features of both HMI and GONG are significantly higher for the extreme values (15–20) \( \times 10^{22} \) Mx. For the HMI test data and GONG data, mean absolute errors are more than 12 times higher at large magnitudes (>1000 \( G^2 - m^{-1} \)) of absolute net current helicity compared to the HMI validation data. The average relative errors for GONG and HMI are comparable at \( \approx 80\% \pm 10\% \), \( 900\% \pm 100\% \), and \( 25\% \pm 2\% \) for total unsigned flux, absolute net current helicity, and mean free energy density, respectively. The high average relative errors imply that the CNN estimates
are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.

are far off from true values, particularly for SHARP features with low true values. SHARP features from ARs that produce at least one major flare (M5 or greater) show a significant drop in average relative errors, at approximately 30% ± 15%, 300% ± 80%, and 16% ± 2%, respectively.

4.2. Time Evolution of the CNN-derived Features on Flaring Active Regions

For understanding AR magnetic field dynamics and improving forecasting of solar storms, it is important that temporal variations of the CNN-estimated SHARP is faithful to the true SHARP. We measure trends in the time evolution of SHARP features of an AR by fitting the observed and the CNN-estimated values with smooth spline curves and calculate numerical time derivatives. Table 5 lists Pearson correlations between the time derivative of the true and the CNN-estimated values of total unsigned flux, absolute net current helicity, and mean free energy density. The time derivative is obtained for the true and the CNN-estimated features of each AR after fitting the respective time series to a cubic spline. For reference, the Pearson correlations between the spline fit values of the true and the CNN-estimated features are also shown. Note that these are consistent with the values in Table 2.
from ARs that produce at least one M5 or greater flare are included in Appendix A. The CNN estimation of SHARP features on flaring ARs is thus useful for understanding AR magnetic field evolution leading to particularly violent solar storms in the past. The Halloween storms of 2003 October produced extreme flares from NOAA AR 10486 of magnitudes X17.0, X10.0, and the largest recorded flare X28.0 (Pulkkinen et al. 2005). The magnetic field evolution leading to these extreme flares was characterized by rotation of a major positive polarity of the delta sunspot as shown in the top panel of Figure 6 (Zhang et al. 2008). Without the knowledge of vector magnetic fields, free energy and current helicity during these storms are previously modeled based on the magnetic virial theorem (Metcalf et al. 2005; Régnier & Priest 2007), linear/nonlinear force-free field extrapolation (Régnier & Priest 2007), and a Minimum Current Corona model (Kazachenko et al. 2010). We obtain model-free and purely data-driven CNN estimates of total unsigned flux, absolute net current helicity, and mean free energy density during these storms using LOS magnetograms. However, the HMI observations are not available for this period. We therefore use the CNN trained with HMI magnetograms to process LOS observations from MDI during the Halloween storms to estimate time evolution of total unsigned flux, absolute net current helicity, and mean free energy density. Flare X28.0 is excluded because it occurred outside 45° of the central meridian. In particular, the CNN-estimated absolute net current helicity of NOAA AR 10486 rises continuously by 25% between the X1.2 flare and the X17.0 flare corresponding to the observed sunspot rotation. A similar gradual rise of a modeled helicity flux by 50% between the X1.2 and X17.0 flares has been reported (Kazachenko et al. 2010), caused primarily by helicity injection from the rotation of the sunspot. The CNN estimates show that the absolute net current helicity stays high leading to the X10.0 flare and falls thereafter. The CNN-estimated mean free energy density also rises leading to the X17.0 and X10.0 flares. Note that these CNN-estimated values from MDI magnetograms are not expected to be corrected for the instrument cross-calibration between the MDI and HMI since the CNN is trained with only HMI magnetograms. Table 8 in Appendix B lists the Pearson and Spearman correlations between the true values and the CNN-estimated values using MDI LOS magnetograms, during the overlap period of MDI and HMI. These correlation values are significantly lower compared to those estimated from HMI magnetograms (Tables 2 and 3). Therefore, a rigorous estimation first requires standardization of MDI and HMI magnetograms (e.g., with other approaches such as super-resolution; Munoz-Jaramillo et al. 2022). We also used the GONG magnetograms to estimate the vector field features during the storms using the CNN trained with GONG (see Appendix C). The values of the vector field features estimated using GONG magnetograms are in the extreme range, as expected during the storms. However, the sensitivity of these estimated values to pre- and post-flare magnetic field variations is lower compared to the features estimated from MDI.

Figure 5. Comparison of true and CNN-estimated SHARP vector field features. Comparisons of time evolution of the CNN-estimated total unsigned flux, absolute net current helicity, and mean free energy density with true values for ARs that produce M5 or greater flares. Only AR observations within ±45° of the central meridian are considered. The left panel shows a typical result (see Figure 9 for all ARs with major flares). The right panel shows an extreme event with the largest flare observed in cycle 24. The gaps correspond to the missing observations and 1σ error bars are shown. The legend in the top left panel applies to all plots. The plots are smoothed with a 6 hr running average.
4.3. Flare Forecasting Using CNN-derived Features

The SHARP features have been extensively used for building flare forecasting models using ML (Bobra & Couvidat 2015; Bobra & Ilonidis 2016; Nishizuka et al. 2017; Chen et al. 2019; Dhuri et al. 2019; Ahmadzadeh et al. 2021). In order to assess the utility of CNN-estimated SHARP for flare forecasting tasks, we compare their flare forecasting performance to the true SHARP. We set up the problem of forecasting M-/X-class flares with 24 hr warning, similarly to Bobra & Couvidat (2015). We use two approaches for the comparison. First, we build Linear Discriminant Analysis (LDA) classification models using one SHARP feature at a time. This allows for direct comparison of the true and the CNN-estimated values of each SHARP feature for flare forecasting. Second, we use all SHARP features together to train a support vector machine (SVM) for flare forecasting. We measure the flare forecasting performance using accuracy, recall, and the True Skill Statistics (TSS) score (Peirce 1884). However, only the latter two are robust to the class imbalance prevalent in the flare forecasting problem (Bobra & Couvidat 2015; Ahmadzadeh et al. 2021) and therefore reliable for comparison. Our definitions of positive and negative classes are identical to the operational approach described in Bobra & Couvidat (2015). In addition, we use the 10 times repeated-holdout validation described in Section 3. Unlike Bobra & Couvidat (2015), we explicitly ensure that the samples from a given AR are not mixed in training and validation sets (Ahmadzadeh et al. 2021). In addition, as mentioned in Section 2, we only consider ARs with maximum area >25 Mm². Both the LDA and SVM are implemented using the scikit-learn library in Python.

Table 6 lists performance metrics for the classification of M-/X-class flares using the LDA of one SHARP feature at a time. The accuracy, recall, and TSS values obtained using each of the CNN-estimated features from HMI and GONG magnetograms are consistent with those of the true SHARP features up to the validation error bars. We note that Schrijver’s R_value (Schrijver 2007) gives the highest TSS values for flare forecasting using individual features.

Table 7 lists the performance metrics for the SVM classification of M-/X-class flares using all SHARP features together. TSS (∼68%) and recall (∼86%) values obtained using an SVM trained with the CNN-estimated features from

![Figure 6](image_url). The 2003 Halloween storms. Top: NOAA AR 10486 MDI magnetograms show the AR dynamics leading to extreme flare events X17.0, X10.0, and X28.0 (outside ±45° of the central meridian, not shown) characterized by rotation of the positive polarity spot (Zhang et al. 2008; Kazachenko et al. 2010), highlighted within the red circle. Positive (white) and negative (black) polarities are saturated at 1000 G. Bottom: the CNN-estimated total unsigned flux, mean free energy density, and absolute net current helicity during these storms showing a systematic rise leading to the flares. The CNN accurately captures helicity injection due to the sunspot rotation showing a significant increase in the mean free energy density and electric current helicity. The gaps correspond to the missing observations, and 1σ error bars are shown. The plots are smoothed with a 6 hr running average.
HMI are consistent with those obtained using the true SHARP. TSS (~62%) and recall (~80%) values from an SVM trained with the CNN-estimated features from GONG are slightly lower. For a comparison, we list TSS (~76%) and recall (~83%) from Bobra & Couvidat (2015) that are higher. The systematically lower TSS of the SVM in forecasting flares when using true SHARP values here as compared with Bobra & Couvidat (2015) is due to exclusion of observations from ARs with maximum area <25 Mn² (all nonflaring) and the explicit restriction that samples from an AR are part of either training or validation sets. Largely consistent performance metrics for flare forecasting with the CNN-estimated SHARP imply that, high relative errors notwithstanding, the CNN-estimated features can be useful for building space weather forecasting tools. This is a consequence of (true) SHARP feature values varying over several orders of magnitudes and thus being significantly different for flaring and nonflaring ARs for forecasting of flares (Dhuri et al. 2019). Accuracy of the CNN-estimated SHARP features may be improved by significantly increasing the resolution of LOS magnetograms from, e.g., GONG, using techniques such as super-resolution (Munoz-Jaramillo et al. 2022). Our method is thus suitable for reconstructing vector field features from historical LOS magnetograms, ultimately useful for reliable space weather forecasting.

### 4.4. Interpreting the CNN

CNNs and, in general, deep learning are extremely efficient at identifying correlations in the data. In this case, the CNN builds a useful model of AR vector magnetic fields from the observed LOS magnetograms. In particular, the CNN-estimated SHARP features may be reliably used to study energy buildup and time evolution of magnetic fields in flaring ARs. Yet it is very challenging to open up the trained network and understand the CNN to uncover the information absorbed. Nevertheless, weights learned by the CNN can shed some light on its working. There are also attribution methods to quantify the contribution of different parts of the input image to the CNN’s output. Here we analyze the weights of the CNN and

| Feature | True SHARP | CNN-HMI | CNN-GONG |
|---------|------------|---------|-----------|
| Total unsigned flux | 87.04 ± 0.36 | 85.01 ± 0.69 | 87.01 ± 0.24 |
| Area | 85.27 ± 0.16 | 85.34 ± 0.18 | 84.32 ± 0.57 |
| Total unsigned vertical current | 88.19 ± 0.27 | 85.51 ± 0.28 | 87.06 ± 0.23 |
| Total unsigned current helicity | 89.81 ± 0.82 | 91.04 ± 0.54 | 88.02 ± 0.20 |
| Total free energy density | 89.77 ± 0.04 | 71.04 ± 0.75 | 86.04 ± 0.74 |
| Total Lorentz force | 87.68 ± 0.47 | 69.81 ± 0.62 | 88.70 ± 0.03 |
| Absolute net current helicity | 91.01 ± 0.27 | 61.87 ± 0.08 | 72.11 ± 0.67 |
| Sum of net current per polarity | 90.96 ± 0.19 | 89.48 ± 0.30 | 87.03 ± 0.29 |
| Mean free energy density | 76.56 ± 0.94 | 86.92 ± 0.14 | 73.94 ± 0.18 |
| Area with shear >45° | 67.15 ± 0.32 | 71.06 ± 0.45 | 76.72 ± 0.20 |
| Log of flux near polarity inversion line | 67.57 ± 0.83 | 97.98 ± 0.16 | 65.83 ± 0.05 |

Note. 1σ standard deviation is shown.
obtain attribution maps for input magnetograms to interpret the trained CNN.

The CNN architecture (Figure 1) comprises a fully convolutional network for processing LOS magnetograms and a fully connected layer of neurons for processing information about the location of ARs on the solar disk. The penultimate concatenation layer comprises a global-average-pooling layer, a global-max-pooling layer that processes the LOS magnetograms, and a fully connected layer that processes the location of the ARs. The global-average-pooling neurons are sensitive to the entire spatial extent of LOS magnetograms, while the global-max-pooling neurons are sensitive to spatially local patterns. The fully connected neurons are sensitive to AR coordinates on the disk. Figure 7 illustrates the distribution of the top weights of each of the three components in the penultimate layers as their contribution to the output of the CNN that estimates total unsigned flux, absolute net current helicity, and mean free energy density. Neurons associated with global average pooling contribute dominantly to the total unsigned flux and mean free energy density, implying that their estimation depends on the consideration of the entire LOS magnetograms. For absolute net current helicity, key contributors are neurons from the global-max-pooling layer, and its estimation is sensitive to spatially local patterns from LOS magnetograms. Without the global-max-pooling layer, absolute net current helicity and related CNN-estimated SHARP features show ∼30% less Pearson correlation with the true values. Weights from neurons related to AR location on the solar disk are ∼0, and thus the CNN estimation does not strongly depend on the AR location. Indeed, the CNN may be trained equally well without the additional input of the AR location. This may be a consequence of considering AR patches only within ±45°, where the projection effects are not significant.

While there are many attribution methods, gradient-based methods such as saliency maps (Simonyan et al. 2013), grad-CAMs (Selvaraju et al. 2017), integrated gradients (IG; Sundararajan et al. 2017), etc., are favored over perturbation-based methods such as occlusion masks (Zeiler & Fergus 2014) because of computational efficiency and higher-resolution attribution maps. IG attribution maps are of the same resolution as the input magnetograms and are thus superior to grad-CAMs obtained from the CNN feature maps. In addition, unlike saliency maps, IG attribution maps are calculated using a reference input image that facilitates assigning a cause for the attribution, e.g., by comparing the magnetic field evolution (Sun et al. 2022). Thus, here we use IG attribution maps to identify pixels, and hence the magnetic field features in the input, that are important for the CNN output. The IG attribution map for a given input image is calculated by integrating gradients in the CNN output along the path from a reference image. Formally,

\[
L^f(x, x_0) = (x - x_0) \times \int_{\alpha=0}^1 \frac{\partial Y^f(x_0 + \alpha \times (x - x_0))}{\partial x} \, d\alpha,
\]

where \( x_0 \) is the reference image and \( Y^f \) is the CNN output for SHARP vector field feature \( f \).

Figure 8 shows contour plots of typical IG attribution maps for a few example magnetograms from flaring ARs (bottom rows). The red/blue contours include regions of net positive/negative contribution toward the CNN output. The IG attribution maps for the three SHARP features—total unsigned

| # Positives | # Negatives |
|-------------|-------------|
| True SHARP  | 0.842 ± 0.030 | 0.856 ± 0.044 | 0.841 ± 0.033 | 0.697 ± 0.045 |
| CNN:HMI     | 0.812 ± 0.028 | 0.869 ± 0.056 | 0.809 ± 0.031 | 0.677 ± 0.046 |
| CNN:GONG    | 0.818 ± 0.031 | 0.801 ± 0.064 | 0.819 ± 0.035 | 0.621 ± 0.056 |
| True SHARP (Bobra & Couvidat 2015) | 0.924 ± 0.007 | 0.832 ± 0.042 | 0.929 ± 0.008 | 0.761 ± 0.039 |

| Accuracy   | Recall(+) | Recall(−) | TSS    |
|-------------|------------|------------|--------|
| True SHARP  | 0.842 ± 0.030 | 0.856 ± 0.044 | 0.841 ± 0.033 | 0.697 ± 0.045 |
| CNN:HMI     | 0.812 ± 0.028 | 0.869 ± 0.056 | 0.809 ± 0.031 | 0.677 ± 0.046 |
| CNN:GONG    | 0.818 ± 0.031 | 0.801 ± 0.064 | 0.819 ± 0.035 | 0.621 ± 0.056 |
| True SHARP (Bobra & Couvidat 2015) | 0.924 ± 0.007 | 0.832 ± 0.042 | 0.929 ± 0.008 | 0.761 ± 0.039 |

Note. Flare forecasting performance of an SVM (Cortes & Vapnik 1995; Hastie et al. 2001) trained using CNN-estimated SHARP features (Table 2). The SVM is trained to forecast M- and X-class flares 24 hr in advance, similarly to Bobra & Couvidat (2015). 1σ standard deviation is shown.
Figure 8. IG attribution maps. The contour plots for the IG attribution maps highlight the regions from magnetograms with the highest attributions for the CNN estimation of SHARP features—total unsigned flux, absolute net current helicity, and mean free energy density. The red/blue contours correspond to the regions with the net positive/negative attributions for the magnetograms in the bottom rows relative to the reference magnetograms in the top rows. The $\Delta$ values are the percentage change in the normalized values of the respective features between the present and the reference observations. The color bar for the magnetograms is saturated with $\pm 500$ G. The IG attribution maps are smoothed with a Gaussian filter of standard deviation 10 pixels before obtaining the contours.
flux, absolute net current helicity, and mean free energy density—are shown separately along with the reference magnetograms (top rows) used. In general, increasing/decreasing positive polarity flux corresponds to net positive/negative attribution. For total unsigned flux, almost all magnetic field regions, even relatively smaller regions with weaker magnetic fields, constitute a positive/negative attribution. In contrast, for absolute net current helicity and mean free energy density, only relatively larger and stronger magnetic field regions constitute a positive/negative attribution. For mean free energy, positive/negative attribution regions typically correspond to the uniformly increasing/decreasing positive flux. In the case of absolute net current helicity, attributions correspond to regions with “mixed” magnetic fields of the positive/negative polarities closely located. The appearance of a spurious magnetic field polarity inversion line (PIL) is a known artifact in the LOS magnetograms whenever the magnetic field inclination relative to the LOS exceeds 90° (Leka et al. 2017). We find that in many cases (e.g., HARPs 407, 3291, and 3311) when the PIL artifact exists for magnetic fields within penumbrae, it wrongly constitutes an important attribution. The misattribution results from the failure of the CNN to learn the PIL artifact (Sun et al. 2022) and, as a consequence, limits the accuracy of the reconstructed the vector field features. Additional examples of IG attribution maps are shown in Appendix D.

5. Discussion

We have thus developed a CNN model for quantifying vector field properties—extensive features such as total unsigned flux and properties depending explicitly on transverse magnetic field component such as free energy density and current helicity—using LOS magnetograms taken from space-based HMI and ground-based GONG instruments. The CNN-estimated features strongly correlate (>90%) with their true measurements from HMI SHARP, particularly for high-resolution LOS magnetograms from HMI. Time evolution of the CNN-estimated features reliably mimics true AR magnetic field evolution, particularly for ARs producing major flares (M5 or greater). Prior to HMI, vector magnetic field observations available from instruments such as Imaging Vector Magnetograph and Hinode/Spectro Polarimeter (Kosugi et al. 2007) have limited spatial and temporal converge. In contrast, near-continuous observations of LOS magnetograms are available since the 1970s from missions such as the Kitt Peak telescope (KP), MDI, and GONG. LOS magnetograms from these instruments vary in their spatial resolutions that are lower than HMI resolution. Nonetheless, these instruments’ observation periods overlap with HMI (KP: 2010–present; MDI: 2010–2011; GONG: 2010–present), and the attendant observations may be used to train or fine-tune the CNN model to estimate SHARP vector field features. We explicitly show that the flare forecasting performance of the CNN-estimated features is comparable to the true SHARP. Therefore, vector fields estimated from past LOS observations of nearly five decades using CNN can provide approximately four times more solar storm data than currently available, useful for building robust statistical models for space weather forecasting using ML. A larger sample size of solar storms also facilitates building ML algorithms based on time series of AR observations that may significantly improve forecasting performance (Dhuri et al. 2019). The CNN-estimated vector fields also provide a new perspective to understand and quantify magnetic field dynamics during the past extreme events such as the 2003 Halloween storms as demonstrated here.

Our CNN estimates are reliable for studies of solar storms, yet there is also a significant scope of improvement. Our estimates of vector field features using HMI magnetograms are consistently more accurate compared to those estimated using lower-resolution GONG magnetograms. Using LOS magnetograms from GONG and other instruments that are explicitly cross-calibrated with HMI LOS magnetograms may significantly improve accuracy of the corresponding vector field instruments. In addition, deep-learning-based techniques for improving the resolution of magnetograms, namely, super-resolution, are being successfully developed (Rahman et al. 2020; Munoz-Jaramillo et al. 2022). Using super-resolved LOS magnetograms as input to the CNN promises to yield more accurate CNN estimates of the vector field features. Our estimates are also based only on the training data from the rising phase of cycle 24. Using new data available from HMI and also from newer instruments, a robust CNN regression is achievable. Extending our method, reasonable data-driven estimates of even the full photospheric vector magnetic field from only LOS magnetograms may be feasible, which opens up a new approach in studying and modeling AR magnetic fields using ML.

S.M.H. acknowledges funding from Department of Atomic Energy grant RTI4002 and the Max-Planck Partner Group program. D.B.D. acknowledges funding from Tamkeen under the NYU Abu Dhabi Research Institute grant G1502. D.B.D. and S.M.H. acknowledge discussions with Mark C. M. Cheung and Marc DeRosa. The authors would also like to thank the anonymous reviewer and the scientific editor Manolis K. Georgoulis for their comments and suggestions that helped improve the clarity of the manuscript. The authors declare that they have no competing interests. D.B.D. and S.M.H. designed the research. D.B.D., S.B., and S.K.M. analyzed data. D.B.D. and S.M.H. interpreted the results. D.B.D. wrote the manuscript, with contributions from S.M.H. HMI LOS and vector magnetograms, MDI LOS magnetograms, and SHARP data are publicly accessible on the JSOC data server at http://jsoc.stanford.edu/, courtesy of the HMI and MDI science teams. The GONG LOS magnetograms are publicly available at https://gong.nso.edu/ and were acquired by GONG instruments operated by NISP/NSO/AURA/NSF with contribution from NOAA.

Appendix A

Time Evolution of the CNN-estimated Features on All ARs Producing Flares M5 or Greater

Figure 9 shows time evolution of the CNN-estimated features on all ARs producing at least one M5 or greater flare.
MDI Correlations

Table 8 shows Pearson and Spearman correlations between the CNN-estimated SHARP features using MDI LOS magnetograms and their true values.

Figure 9. Time evolution of the CNN-estimated vector field features on flare-productive ARs. Comparison of the CNN-estimated values (using HMI (blue) and GONG (black) LOS magnetograms) of total unsigned flux, absolute net current helicity, and mean free energy density with true values (red) calculated from HMI vector magnetograms is shown for HARP 750. Only observations within ±45° are considered. 1σ error bars are shown. The gaps indicate missing observations. The complete figure set (28 images) for ARs that produced at least one M5 or greater flare is available in the online journal. (The complete figure set (28 images) is available.)

Appendix B

MDI Correlations

Table 8 shows Pearson and Spearman correlations between the CNN-estimated SHARP features using MDI LOS magnetograms and their true values.

Table 8

| SHARP Features                                      | Pearson Correlation | Spearman Correlation |
|-----------------------------------------------------|---------------------|----------------------|
| Total unsigned flux                                 | 82.49 ± 21.01       | 68.09 ± 37.73        |
| Area                                                | 85.55 ± 23.82       | 74.59 ± 17.72        |
| Total unsigned vertical current                     | 74.83 ± 45.39       | 64.60 ± 46.31        |
| Total unsigned current helicity                     | 75.22 ± 45.28       | 65.80 ± 47.14        |
| Total free energy density                           | 84.18 ± 12.15       | 76.40 ± 19.94        |
| Total Lorentz force                                 | 89.13 ± 11.33       | 75.37 ± 29.10        |
| Absolute net current helicity                       | 51.62 ± 27.28       | 48.97 ± 23.21        |
| Sum of net current per polarity                     | 42.84 ± 35.65       | 38.69 ± 32.27        |
| Mean free energy density                            | 92.60 ± 04.02       | 89.26 ± 07.80        |
| Area with shear >45°                                | 91.78 ± 03.32       | 89.63 ± 03.32        |
| Flux near polarity inversion line                   | 62.64 ± 14.63       | 59.09 ± 21.23        |

Note. The SHARP features are estimated using the CNN trained with the HMI LOS magnetograms. The AR patches of MDI LOS magnetograms are taken from the publicly available data product Space-Weather MDI Active Region Patch (SMARP; Bobra et al. 2021). SMARP and SHARP data overlap between 2010 May 1 and October 28 (Bobra et al. 2021).
Appendix C
Comparison of the CNN-estimated Features during the 2003 Halloween Storms

Figure 10 shows comparison of the CNN-estimated features, using MDI and GONG LOS magnetograms, during the 2003 Halloween storms.

Figure 10. A comparison of the time evolution of the CNN-estimated vector field features during the 2003 Halloween storms using MDI and GONG LOS magnetograms. The CNN trained with HMI magnetograms is used for the estimation from MDI magnetograms, whereas the CNN trained with GONG magnetograms is used for the estimation from GONG magnetograms. Note that the HMI observations are not available before 2010. The CNN:GONG feature values are generally high compared to the CNN:MDI, showing little variation throughout the storms. The CNN:MDI features appear to capture the variation of these features during the storms expected from the theoretical modeling (e.g., Kazachenko et al. 2010). The 1σ errors are shown. The gaps indicate missing observations. The legend in the rightmost panel applies to all panels.
Appendix D
IG Attribution Maps

Figure 11 shows additional examples of IG attribution maps.

Figure 11. Additional examples of IG attribution maps (Figure 8).
References

Ahmadzadeh, A., Aydin, B., Georgoulis, M. K., et al. 2021, ApJS, 254, 23
Bhattacharjee, S., Alshehhi, R., Dhuri, D. B., & Hanasoge, S. M. 2020, ApJ, 898, 98
Bobra, M. G., & Couvidat, S. 2015, ApJ, 798, 135
Bobra, M. G., & Ilonidis, S. 2016, ApJ, 821, 127
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Bobra, M. G., Wright, P. J., Sun, X., & Turmon, M. J. 2021, ApJS, 256, 26
Boteler, D. H. 2019, SpWea, 17, 1427
Bottou, L. 1991, in Proc. Neuro-Nimes 91 (Nimes: EC2), 687
Chen, Y., Manchester, W. B., Hero, A. O., et al. 2019, SpWea, 17, 1404
Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273
Crown, M. D. 2012, SpWea, 10, S06006
Dhuri, D. B., Hanasoge, S. M., & Cheung, M. C. M. 2019, PNAS, 116, 11141
Eastwood, J. P., Bifflis, E., Hapgood, M. A., et al. 2017, RiskA, 37, 206
Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (Cambridge, MA: MIT Press)
Han, J., & Moraga, C. 1995, in From Natural to Artificial Neural Computation, ed. J. Mira & F. Sandoval (Berlin: Springer), 195
Hastie, T., Friedman, J., & Tibshirani, R. 2001, The Elements of Statistical Learning (New York: Springer), doi:10.1007/978-0-387-21606-5
Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483
Kazachenko, M. D., Canfield, R. C., Longcope, D. W., & Qiu, J. 2010, ApJ, 722, 1539
Kosugi, T., Matsuzaki, K., Sakao, T., et al. 2007, SoPh, 243, 3
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Leka, K. D., & Barnes, G. 2007, ApJ, 656, 1173
Leka, K. D., Barnes, G., & Wagner, E. L. 2017, SoPh, 292, 36
Livingston, W. C., Harvey, J., Pierce, A. K., et al. 1976, ApOpt, 15, 33
McIntosh, P. S. 1990, SoPh, 125, 251
Metcalf, T. R., Leka, K. D., & Mickey, D. L. 2005, ApJL, 623, L53
Munoz-Jaramillo, A., Jungbluth, A., Guitaux, X., et al. 2022, Natur. Portfolio, in press
Nishizuka, N., Sugira, K., Kubo, Y., et al. 2017, ApJ, 835, 156
Peirce, C. S. 1884, Sci, ns-4, 453
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, SoPh, 275, 3
Pulkkinen, A., Lindahl, S., Viljanen, A., & Pirjola, R. 2005, SpWea, 3, S08C03
Rahman, S., Moon, Y.-J., Park, E., et al. 2020, ApJL, 897, L32
Régnier, S., & Priest, E. R. 2007, ApJL, 669, L53
Scherrer, P. H., Bogart, R. S., Bush, R. I., et al. 1995, SoPh, 162, 129
Schrijver, C. J. 2007, ApJL, 655, L117
Selvaraju, R. R., Cogswell, M., Das, A., et al. 2017, in 2017 IEEE Int. Conf. on Computer Vision (ICCV) (Piscataway, NJ: IEEE), 618
Shibata, K., & Magara, T. 2011, LRSP, 8, 6
Simonyan, K., Vedaldi, A., & Zisserman, A. 2013, arXiv:1312.6034
Stenflo, J. O. 2013, A&ARv, 21, 66
Su, Y., Veronig, A. M., Holman, G. D., et al. 2013, NatPh, 9, 489
Sun, Z., Bobra, M. G., Wang, X., et al. 2022, ApJ, 931, 163
Sundararajan, M., Taly, A., & Yan, Q. 2017, PMLR, 70, 3328
Szegedy, C., Liu, W., Jia, Y., et al. 2015, in 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ: IEEE), 1
Zeiler, M., & Fergus, R. 2014, in Computer Vision–ECCV 2014, Lecture Notes in Computer Science, ed. D. Fleet et al. Vol. 8689 (Berlin: Springer), 818
Zhang, Y., Liu, J., & Zhang, H. 2008, SoPh, 247, 39