The Astrophysical Journal, 952:111 (8pp), 2023 August 1 https://doi.org/10.3847/1538-4357/acdeea

Prediction Capability of Geomagnetic Events from Solar Wind Data Using Neural Networks

Daniele Telloni[1,4], Maurizio Lo Schiavo[2,4], Enrico Magli[2], Silvano Fineschi[1], Sabrina Guastavino[3], Gianalfredo Nicolini[1], Roberto Susino[1], Silvio Giordano[1], Francesco Amadori[1], Valentina Candiani[3], Anna Maria Massone[3], and Michele Piana[3]
[1] National Institute for Astrophysics, Astrophysical Observatory of Torino, Via Osservatorio 20, I-10025 Pino Torinese, Italy; daniele.telloni@inaf.it
[2] Politecnico of Turin, Department of Electronics and Telecommunications, Corso Duca degli Abruzzi 24, I-10129 Torino, Italy
[3] University of Genoa, Department of Mathematics, Via Dodecaneso 35, I-16146 Genoa, Italy
Received 2023 May 8; revised 2023 May 31; accepted 2023 June 1; published 2023 July 21

Abstract

Multiple neural network architectures, with different structural composition and complexity, are implemented in this study with the aim of providing multi-hour-ahead warnings of severe geomagnetic disturbances, based on in situ measurements of the solar wind plasma and magnetic field acquired at the Lagrangian point L1. First, a statistical analysis of the interplanetary data was performed to point out which are the most relevant parameters to be provided as input to the neural networks, and a preprocessing of the data set was implemented to face its heavy imbalance (the Earth's magnetosphere is in fact mostly at rest). Then, neural networks were tested to evaluate their performance. It turned out that, in a binary classification problem, recurrent approaches are best at predicting critical events both 1 and 8 hr in advance, achieving a balanced accuracy of 94% and 70%, respectively. Finally, in an attempt at multistep prediction of the criticality of future geomagnetic events from 1–8 hr ahead, more complex neural networks, built by merging the different types of basic convolutional and recurrent architectures, have been shown to outperform single-step and state-of-the-art approaches with a balanced accuracy of at least 70%. Interestingly, the accuracy peaks at 4 hr, corresponding to the waiting time between the detection of solar drivers at L1 and the onset of the geomagnetic storm (as previously obtained by statistical investigations), suggesting that on average this is the time the Earth's magnetosphere takes to react to the solar event.

Unified Astronomy Thesaurus concepts: Solar coronal mass ejections (310); Neural networks (1933); Heliosphere (711); Solar-terrestrial interactions (1473); Solar wind (1534); Solar storm (1526); Geomagnetic fields (646)

1. Introduction

Space weather is the science of investigating how solar–terrestrial interaction affects the geospace environment. More specifically, it concerns tracking and predicting severe solar-driven disturbances that can pose serious hazards to ground-based and in-orbit human activities and technological assets. Among these, the most energetic and potentially harmful are arguably the interplanetary counterparts of coronal mass ejections (CMEs; Klein & Burlaga 1982; Webb & Howard 2012), large expulsions of magnetized plasma from the outermost layer of the Sun's atmosphere, the corona, into interplanetary space. Although they are known to be launched by large-scale explosive rearrangements of the solar magnetic field via magnetic reconnection processes, the forecasting of their onset is, to date, beyond capability. At the interface between high- and low-speed solar wind streams, regions of intense magnetofluid compression are formed. As such, they also induce geomagnetic disturbances, albeit of lesser severity than those caused by CMEs. Being corotating interplanetary structures with the Sun, they are generally referred to as corotating interaction regions (CIRs; Gosling & Pizzo 1999). The geomagnetic activity associated with Earth's passage of CMEs and CIRs is reported, e.g., in Gosling et al. (1991) and Richardson et al. (2006), respectively. Another key area pertaining to space weather science, in addition to the prediction of geoeffective solar events, is the forecast of the intensity of geomagnetic storms induced by solar activity (e.g., Telloni et al. 2020) and the duration of the associated recovery phase (Telloni et al. 2021).

Recently, the heliophysics community has been increasingly relying on the use of machine-learning (ML) techniques for space weather science (Camporeale 2019), with the most evident opportunity to develop physics-based and data-driven prediction models that can operate in real time. Most of the effort has been devoted so far to forecasting the onset of CMEs and their time of arrival to Earth, using remote observations of the Sun and its corona. Noteworthy in this regard are the studies by Bobra & Ilonidis (2016), Sudar et al. (2016), Liu et al. (2018), Liu et al. (2020), and Guastavino et al. (2023). On the other hand, the potentialities of ML approaches to in situ measurements acquired by orbiting spacecraft at the L1 Lagrangian point for predictive purposes have been definitely less explored. Some works aim to classify interplanetary structures (such as CMEs, CIRs, high- and low-speed streams, or heliospheric current sheet crossings; Camporeale et al. 2017; Li et al. 2020; Roberts et al. 2020), others to forecast the solar wind plasma and magnetic field profiles at L1 from remote-sensing (Upendran et al. 2020; Raju & Das 2021; Yang & Shen 2021), cosmic-ray (Sabbatini & Grimani 2023), or even geomagnetic (Kataoka & Nakano 2021) data, and that by Reiss et al. (2021) to predict the southward component of the CME magnetic field (a sine qua non condition for magnetic

[4] These authors contributed equally to this work.

reconnection with the Earth's northward magnetic field to occur). In spite of their undoubted impact in space-weather-related applications, however, the above studies are not concerned with predicting the geoeffectiveness of solar structures or estimating their degree of dangerousness. Some attempts have been made at trying to predict geomagnetic activity from in situ solar wind measurements, but either the results are far from being satisfactory or somewhat limited by the deployment of a single class of ML architecture. Bala et al. (2015) employed neural network (NN) algorithms to infer the degree of storminess that the CME ejected on 2012 July 23 would have had if it had hit Earth. Keesee et al. (2020) applied two different NNs to study the relationship between solar drivers at L1 and disturbances in the Earth's magnetosphere. As performance testing on two geomagnetic storms has shown, the algorithms developed by the authors are unable to predict either the severity of disturbances or their duration. Some time forecasting ability emerges, but validation metrics revealed it to be barely higher than random or constant predictions. Instead, the study by Kitajima et al. (2022) reports only that solar wind speed first and density second play key roles in geomagnetic disturbances once the interplanetary magnetic field is directed southward (and thus favorable to magnetic reconnection). This result is expected, however, because higher velocity and/or density implies greater solar plasma kinetic energy being transferred to the Earth's magnetic field on the sunward side during solar-wind-magnetosphere coupling. Since velocity appears squared in the energy expression, clearly it is the most important parameter in this respect. Definitely more interesting studies are presented in Wu & Lundstedt (1997), Bala & Reiff (2012), Lazzús et al. (2017), Gruet et al. (2018), Laperre et al. (2020), Collado-Villaverde et al. (2021), Park et al. (2021), Tasistro-Hart et al. (2021), Wintoft & Wik (2021), and Singh (2022), who provide predictive models for geomagnetic indices based on interplanetary measurements acquired at L1 and relying on different ML approaches. However, most of these works use only one class of NNs (the recurrent models), without exploring how forecasting performances may depend on the particular architecture adopted; others (Park et al. 2021; Singh 2022) employ feedforward NNs that, as discussed below, are not suitable for processing sequential data seeking temporal correlations to be exploited for predictive purposes. Finally, Collado-Villaverde et al. (2021) use only interplanetary magnetic field data as input parameters for the NN, while not considering plasma measurements, which instead (as shown in this paper and also in Li et al. 2022) turn out to be crucial for proper network learning. As a conclusion, there is currently no study in the literature based on in situ solar wind data that aims to evaluate and compare the performance of the most common and the most complex NNs in predicting geomagnetic events, while providing multiple-hour-ahead forecast of the geomagnetic indices. To complete these works, the present study was initiated.

As mentioned just above, in situ solar wind measurements are only poorly exploited for space weather purposes (in contrast to coronagraphic observations) even in the context of ML applications. The reason stems from the challenging local identification of the CMEs (to the extent of resorting to the use of ML techniques for their automatic detection; Nguyen et al. 2019), which in fact prevents an early warning thereof, and in turn, of the major geomagnetic storms. The work by Telloni et al. (2019) represents a first attempt to address this significant deficiency. In fact, the authors have developed a novel algorithm for the proper detection of CMEs at L1 by exploiting their distinguishing feature, namely, the fact that they carry a helical structure, known as flux rope, which can be uniquely revealed by measuring the degree of twisting of the interplanetary magnetic field lines, i.e., the magnetic helicity (Matthaeus & Goldstein 1982; Telloni et al. 2012). By coupling the so-based identification of CMEs with the estimate of their energy content, Telloni et al. (2019) provided a useful tool for forecasting the geoeffectiveness of CMEs. It turned out that the CME-related geomagnetic disturbances can thus be alerted (with an efficiency of 86% for the weakest and 100% for the most severe) with an average warning of 4 hr, ranging between 2 and 8 hr in 98% of instances. It should be emphasized that the 4 hr waiting time between the detection of the CME and the onset of the geomagnetic storm comprises two contributions: the CME flight time from L1 to Earth and the time for the Earth's magnetosphere to respond to the solar driver, i.e., to intensify the ring currents, measured by the SYMmetric perturbations in the Horizontal component of the ground magnetic field vector (SYM-H index; Iyemori 1990).

This paper aims at extending the study by Telloni et al. (2019) by implementing seven NNs, based on different architectures and built with different complexities, for the forecasting of geomagnetic events driven by CMEs. Indeed, the deployment of ML techniques poses a valuable opportunity to validate the statistical results reported in the previous analysis and to refine the estimation of the hazard level of CMEs. Leveraging a large data set of local solar wind measurements and ground-based geomagnetic indices collected over a 15 yr interval, the goal is to select the most suitable architecture for developing an ML-based predictive tool capable of timely forecasting future onset of any CME-induced disturbances, as well as multi-hour advance warning in case critical events may prove harmful to Earth and human activities. Different prediction tasks are proposed here, tackled from the ML perspective as classification problems, which disclose information about the degree of criticality of CMEs and their time duration. This paper is organized as follows: Section 2 presents the NN architectures adopted in the present work, while Section 3 illustrates the predictive capabilities of the developed algorithms along with future perspectives.

2. NNs

The first crucial steps in every ML algorithm implementation are the problem definition, data collection, and data preprocessing. The aim of the NN developed for the present analysis is an early forecasting of when the geomagnetic index SYM-H decreases below -50 nT. This is the lowest commonly accepted threshold for a geomagnetic storm to occur (Cander & Mihajlovic 1998). The SYM-H index is thus referred to by the algorithm as the target variable. Basing the learning process upon past solar wind data that caused SYM-H < -50 nT events, the NN is designed to predict future geomagnetic disturbances. This defines the formulation of the to-be-performed study, that is, the problem definition.

2.1. Data Collection and Preprocessing

Training, validating, and testing any NN requires a huge amount of data. The larger the data set, the better the tuning of the NN. For the present work, the data set consists of

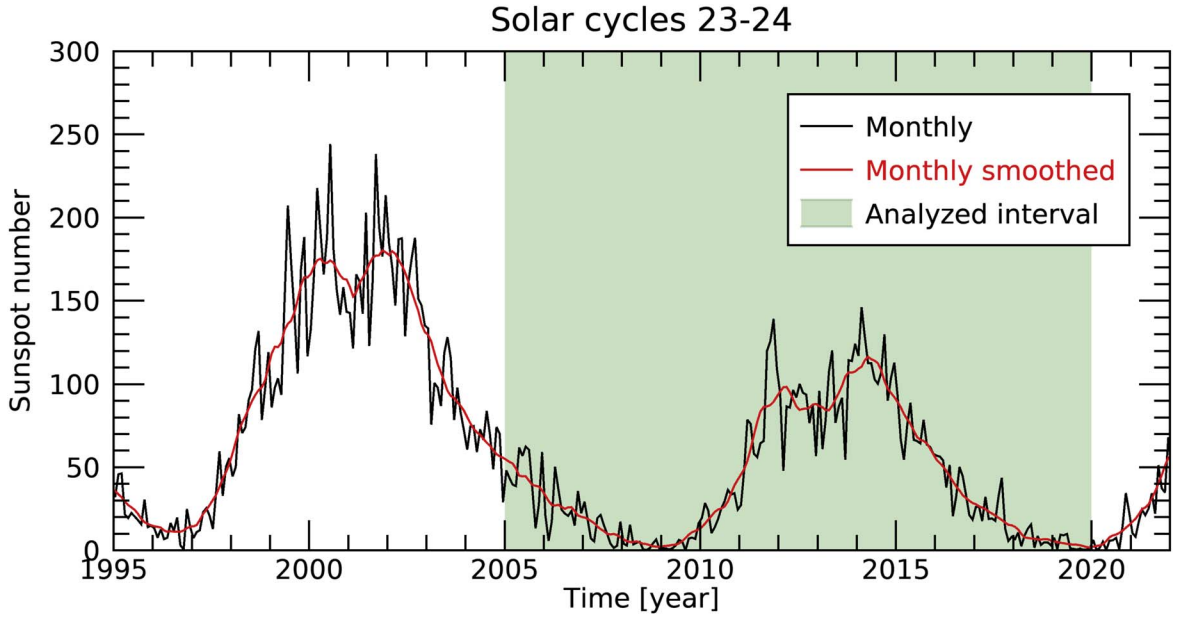

Figure 1. Time profile of monthly (black) and 13 month smoothed (red) averages of the sunspot number for solar cycles 23 and 24. The green shaded area indicates the period of interest for this work.

measurements gathered over a 15 yr period from 2005–2019, spanning from the descending phase of solar cycle 23 to the minimum of solar cycle 25, that is, over more than one solar cycle. The period of interest is shown in the green shaded area of Figure 1, which overall spans from 1995–2022 and illustrates the monthly sunspot number, a common proxy for solar cycle activity.

Aiming to investigate the causal relationship between solar wind and geomagnetic activity, the Operating Missions as a Node on the Internet (OMNI; King & Papitashvili 2005) database is conveniently used. This includes interplanetary data (mainly plasma and magnetic field measurements) acquired at L1 from different spacecraft (such as Wind, the Advanced Composition Explorer, and the Deep Space Climate ObserVatoRy), combined and resampled at 1 minute resolution, and time shifted as if they had been acquired at the nose of Earth's bow shock (i.e., to the acquisition time is added the time required for the solar wind plasma parcel to travel the distance L1—Earth). The OMNI data set also comprises geomagnetic data including the 1 minute resolution SYM-H index. Among the interplanetary parameters in the OMNI database, only a subset was selected for the ML-based multivariate analysis as the features most influencing the SYM-H target value. These are the intensity B and southward-directed component B_z of the magnetic field vector, and the solar wind speed U, density n, and temperature T. These in fact turn out to be the most related to SYM-H, as evidenced by the last row (column) in Table 1, which overall reports the correlation matrix, i.e., the absolute Pearson correlation coefficients between the selected solar wind parameters and the magnetospheric index: values higher than 0.2 suggest a relevant contribution of these features to geomagnetic activity.

This is not surprising, however. As a matter of fact, the main characteristics of interplanetary CMEs are higher magnetic field strengths and lower densities and temperatures (Klein & Burlaga 1982). The correlation of B, n, and T with SYM-H, which emerges from the background represented by the mostly unperturbed solar wind, thus only points to the CMEs as the major drivers of geomagnetic storms. In addition, as mentioned above, it is necessary for the interplanetary magnetic field to be directed southward to reconnect with the Earth's magnetic field, while a higher plasma velocity (and to a lesser extent, a denser plasma) implies a higher energy content, and in turn, a greater release of energy to the near-Earth space environment. Therefore, these two features are also expected to be correlated with SYM-H. Notwithstanding, the highlighted relationship of these features with the targeted variable was imperative in setting up the NN in order not to consider quantities unrelated to geomagnetic activity that could adversely affect the model performance. Instead, the selection of only those features suggesting some degree of dependence with SYM-H (i.e., a potential forecasting support) and thus likely to be informative for the prediction of geoeffective events ensures proper data collection. Incidentally, it is also worth noting the good correlation between the considered solar wind parameters (with the exception of B_z), and especially the well-known $T-U$ statistical relationship (e.g., Burlaga & Ogilvie 1970; Lopez & Freeman 1986), which can be interpreted as emerging from the symmetry of the solar wind transport equations (Matthaeus et al. 2006).

Table 1
Correlation Matrix

	B [nT]	B_z [nT]	U [km s^{-1}]	n [cm^{-3}]	T [K]	SYM-H
B [nT]	1.00	0.01	0.16	0.29	0.30	0.27
B_z [nT]	0.01	1.00	<0.01	<0.01	<0.01	0.22
U [km s^{-1}]	0.16	<0.01	1.00	0.36	0.70	0.47
n [cm^{-3}]	0.29	<0.01	0.36	1.00	0.15	0.30
T [K]	0.30	<0.01	0.70	0.15	1.00	0.34
SYM-H	0.27	0.22	0.47	0.30	0.34	1.00

While a high data volume is necessary, it is not sufficient for proper algorithm training. For an NN to be trained properly, the sample of the target variable, here SYM-H, must be balanced between critical (SYM-H < -50 nT) and noncritical (SYM-H

> -50 nT) occurrences. However, in the present case, only 2% of the time the Earth's magnetosphere was out of equilibrium condition, that is, the SYM-H was less than -50 nT. This leads to a binary classification problem with a highly imbalanced data set. From the ML point of view, the class-imbalance problem can be tackled by using suitable weighted loss functions (Marchetti et al. 2022) or data augmentation. Appropriate data preprocessing aimed at a (partial) rebalancing of the two classes is therefore used in this study. This includes the deployment of different techniques aimed at data readiness for the prediction step. Specifically, after data cleaning (i.e., recovering missing data through, e.g., interpolation) and windowing (i.e., framing the time series into sliding subsections to increase the dimension shape of the data set), augmentation strategies have been implemented to cope with the imbalance of the present data set, basically undersampling the noncritical class and oversampling the critical one. In particular, jittering, scaling, time warping, and pattern-mixing procedures have been equally employed in the preprocessing step of solar wind data analysis. First, white noise (jitter) has been added to the training data set to aid optimization and reduce generalization error. Scaling all the variables under study on the same scale (i.e., standardization) removes the potential adverse influence of different feature variances in the training stage, reducing the convergence time of the learning algorithm, and thus, ensuring computational efficiency. Dealing with plasma and magnetic field variables whose temporal evolution may be very different during geoeffective solar events may compromise the disclosure of similarities between various time series, which is vital for increasing the predictive capabilities of NNs. Time-warping algorithms have been therefore adopted. Basically, these distort the input signal by nonlinearly warping its trend in the time domain, shrinking, or stretching a given time window, so as to equalize the temporal variability of different features. Finally, the more complex pattern-mixing-based Synthetic Minority Oversampling TEchnique (SMOTE) has been exploited to artificially generate new samples belonging to the underrepresented critical class, by simply averaging two existing patterns. The combined use of the above techniques increased the critical samples (SYM-H < -50 nT) up to 32% of the entire data set. The interested reader is referred to Iwana & Uchida (2021) for much more detail on these and many other data augmentation techniques. 70% of the cleaned and windowed data set is used to train the NNs, 20% to validate the quality of the training phase, while the remaining 10% is employed to test the architectures in forecasting future geomagnetic events. Although this division is commonly used in the literature, it was preliminarily verified during the set up of the NNs that the performance of the algorithms does not change appreciably if the proportion between the different stages is slightly changed by a few percent. It is also worth noting that data augmentation has been carried out only to the part of the data set involved in model training and validation (i.e., 90% of the whole data set), while leaving the test set unprocessed. Because CMEs (and in turn, the main phase of the induced geomagnetic disturbances) are events with an average duration at L1 of about 1 day (Klein & Burlaga 1982), the length of the time window considered in the analysis is 24 samples, equally spaced in time by 1 hr so as to represent a daily trend. The total number of time windows thus generated from the OMNI data set turns out to be 7,888,296.

2.2. NN Architectures

Once achieved a better balance between critical and noncritical classes, the NN can then be trained. The best architecture to be employed depends on the aim of the study. Recurrent neural networks (RNNs) are expected to be best suited for predictive purposes, as they deal with time series and can be trained to retain relevant past information to predict future events. Specifically, RNNs are a variant of ordinary feedforward neural networks (FNNs), generally used for simple classification problems. Unlike feedforward architectures, which are designed to handle mutually independent data inputs, RNNs allow the processing of data whose state at time instant t_i depends to some extent (which the NN aims to assess) on the state at time instant t_{i-1} (assuming the time series is not random and governed by some physical process such as the interplanetary/geomagnetic ones studied here). From a mathematical point of view, the predictive capabilities of RNNs are achieved by allowing the model (unlike FNNs, in which information moves in one direction from input and output nodes) to loop back the data flow: the output of a particular layer is fed back to previous steps using it as input to improve predictions. The feedback loop thus allows RNNs to relate inputs to those already processed, in order to assess whether any dependence exists between them and thereby more accurately predict the output of the layer. For the sake of completeness, convolutional neural networks (CNNs), aimed mainly at object detection and image classification, are also worth mentioning. Although recurrent approaches seem better tailored to the nature of the problem under study, in which temporal dependence between consecutive measurements is a key issue for prediction purposes, the implementation of a heterogeneous set of NNs may lead to greater benefits, as the predictive nature of the problem is ultimately tackled as a binary classification challenge (prediction of critical, SYM-H < -50 nT, or noncritical, SYM-H > -50 nT, events). Furthermore, the continuous evolution and refinement of NNs prevent the total supremacy of one architecture over the others for a specific use case. Therefore, a diversified set of networks including FNNs, CNNs, RNNs, and some hybrid architectures (designed to exploit the pros and limit the cons of basic networks, by mixing them or implementing additional arrangements), were considered in this paper. Without discussing in detail, which is beyond the scope of this study (the reader is referred to Aggarwal 2018), for an exhaustive description of the different network architectures), seven different NNs are employed in the present analysis. These are listed below, with a brief description of both their layered structural composition and the reasons for some of the solutions adopted to improve forecasting performance. In setting up each architecture, the same input layer was adopted, as well as the same output shape, i.e., binary classification between critical or noncritical events, while the seven NNs differ in complexity and equipped solutions.

1. The linear single-layer architecture (Bishop 1995) is the simplest FNN and is used as a baseline reference, although its lack of complexity may result in poor predictive ability.
2. The multilayer perceptron (MLP) feedforward architecture (Hornik et al. 1989) is a natural development of linear models: a more complex multilayer structure can improve performance.

3. The CNN (LeCun et al. 2015) is mostly designated to reveal spatial relationships in images and thus find specific patterns in object recognition; notwithstanding, it has also recently been employed on time series for predictive purposes (Raju & Das 2021).
4. The long short-term memory (LSTM) model (Hochreiter & Schmidhuber 1997) is the most widely adopted type of RNN, which solves the well-known short-term memory problem of basic recurrent architectures, which essentially prevents them from propagating information from near the output back to the layers closest to the input. In addition to the above NNs, which belong to the basic network categories, three more complex models, which combine the basic architectures, are also considered.
5. The deep CNN (Goodfellow et al. 1996) is an architecture obtained by replicating 4 times the basic CNN model. In fact, stacking multiple CNN layers can increase the ability of the network to reveal correlations between input data.
6. The LSTM-FCN model (Karim et al. 1998) was designed by merging the recurrent and convolutional extractors of the aforementioned LSTM and deep CNN architectures: combining the outputs extracted with different criteria may indeed enhance the NN efficiency.
7. The deep CNN with skip connection (deep CNN_{skip}; He et al. 2016) is the deep CNN equipped with a skip connection that feeds the output of the network multilayer extraction block with input data. This last model is intended to combine the capabilities of convolutional networks to detect characteristic patterns in time series and the more predictive ones provided by recurrent networks.

Despite differing in their structural composition, the seven NNs employed in this study share common working principles. Basically, each architecture consists of a module (possibly multilayered) that extracts high-level characteristics from the input data. This extraction then feeds one or more fully connected layers aimed at learning potential relationships between these features through the minimization of an a priori chosen loss function. Finally, a classification layer receives the information flow and maps it into the two binary (critical or noncritical) classes, assigning the corresponding probability. As for the extraction phase, it is worth noting that it can be performed jointly, i.e., processing all the features as a whole, or disjointly, i.e., equipping the NN with several extractor modules, as many as the number of features, working in parallel, each in charge of extracting the key characteristics of a single feature. The latter solution can be beneficial for investigating processes in which different features conflict, and simpler architectures generally benefit from it. Indeed, it has been tested that the simpler linear, MLP, CNN, and RNN models yield better results when operating in disjoint mode; on the other hand, the more complex deep CNN, LSTM-FCN, and deep CNN_{skip} architectures perform slightly better with the joint approach. Table 2 gives further details on the settings of the considered NNs.

2.3. Evaluation Metrics

The performance evaluation of the proposed neural architectures is the final stage of the ML process and is carried out on the 10% test set. In a binary classification problem such as the one under study, it thus involves quantifying the percentage of times the algorithm correctly predicted a critical event (SYM-H < -50 nT) and equally a noncritical event (SYM-H > -50 nT). It is indeed evident that a reliable geomagnetic activity prediction tool should not provide false alerts either. The Precision (P), Sensitivity (S), F1-score, and Balanced Accuracy (BA) metrics (particularly useful when dealing with imbalanced data sets) were thus used to judge and measure the efficiency of the trained NNs. Defining T_C and T_{NC} (F_C and F_{NC}) as the correct (incorrect), i.e., True (False), predictions for the two Critical (C) and Noncritical (NC) classes, the precision metric for the critical class, P_C, is defined as

$$P_C = \frac{T_C}{T_C + F_C} \tag{1}$$

and quantifies the number of correct critical predictions out of all predictions (correct or not) of critical events. Similarly, S_C, the sensitivity for the critical class, is the proportion of critical events correctly predicted and is given by

$$S_C = \frac{T_C}{T_C + F_{NC}}. \tag{2}$$

When assessing these two metrics, it is worth mentioning that precision and sensitivity come at the cost of each other. For example, increasing sensitivity means missing no critical events, which allows many false positives F_C to introduce, thus reducing precision. The F1 score is the harmonic mean of P_C and S_C,

$$F1_C = \frac{2P_C S_C}{P_C + S_C}, \tag{3}$$

and keeps the balance between precision and sensitivity. Mirror formulas hold for P, S, and F1-score metrics related to the noncritical class. In particular, S_{NC}, i.e., the sensitivity in forecasting noncritical events, is known as specificity. The arithmetic mean between sensitivity and specificity yields the balanced accuracy estimate of the ML model,

$$BA = \frac{S_C + S_{NC}}{2} = \frac{1}{2}\left(\frac{T_C}{T_C + F_{NC}} + \frac{T_{NC}}{T_{NC} + F_C}\right), \tag{4}$$

which gives equal relevance to the two classes, despite their imbalance.

3. Prediction Capabilities

The seven NN architectures described above (linear, MLP, CNN, LSTM, Deep CNN, LSTM-FCN, and deep CNN_{skip}) have been first tested in forecasting critical geomagnetic disturbances (i.e., sustained periods with SYM-H < -50 nT) 1 and 8 hr in advance, in a single-step approach. The choice stems from the observational result obtained by Telloni et al. (2019) showing that the delay between the detection of any CME and the onset of the induced geomagnetic storm is, in 98% of cases, at most 8 hr. The performance metrics results for these binary classifications are reported in Tables 3 and 4, respectively.

It appears readily evident when attempting to predict events in the near future (1 hr ahead) that the simplest models (even the linear one) provide the best results. This is quite expected since the shorter the time horizon of the forecast, the less complexity is required by the model to predict an imminent

Table 2
Settings of the Employed Network Architectures

Network	Extraction Layers	Input Layer Nodes	Hidden Layer Nodes	Output Layer Nodes
Linear	0	0	144	2
MLP	2	96 + 80	80	2
CNN	1	720	64	2
LSTM	1	72	64	2
Deep CNN	4	9216 × 4	12	2
LSTM-FCN	5	1536 × 4 + 120	64	2
Deep CNN_{Skip}	4	9216 × 4	12	2
	Activation[a] and Loss Function	Number of Epochs	Batch Size and Learning Rate[b]	Optimizer
Linear	ReLU[c] and cross-entropy	100	128 and 0.0001	Adam
MLP	ReLU and cross-entropy	100	128 and 0.00001	Adam
CNN	ReLU and cross-entropy	100	256 and 0.0001	Adam
LSTM	ReLU and cross-entropy	100	256 and 0.0001	Adam
Deep CNN	ReLU and cross-entropy	100	256 and 0.001	Adam
LSTM-FCN	ReLU and cross-entropy	100	128 and 0.00001	Adam
Deep CNN_{skip}	ReLU and cross-entropy	100	64 and 0.00001	Adam

Notes.
[a] For all layers except output ones, for which the SoftMax activation function was used.
[b] The best model configurations were obtained after a series of tests involving different combinations of batch size and learning rate.
[c] Rectified linear unit.

Table 3
Performance Metrics Values of the Seven NNs Considered in the Analysis, for 1 hr Ahead Warning of Critical Geomagnetic Events

Network	P_C	S_C	$F1_C$	BA
Linear	0.641	0.857	0.734	0.927
MLP	0.652	0.819	0.726	0.908
CNN	0.648	0.843	0.733	0.920
LSTM	0.620	0.876	0.736	0.936
Deep CNN	0.628	0.772	0.693	0.885
LSTM-FCN	0.614	0.806	0.697	0.902
Deep CNN_{skip}	0.628	0.823	0.712	0.910

Note. The highest-performing architecture is in bold.

Table 4
Same as Table 3, but for a Forecast 8 hr in Advance and Only the Five Most Efficient Architectures

Network	P_C	S_C	$F1_C$	BA
Linear	0.550	0.311	0.397	0.526
CNN	0.397	0.301	0.342	0.649
LSTM	0.588	0.324	0.418	0.662
LSTM-FCN	0.694	0.391	0.500	0.695
Deep CNN_{skip}	0.708	0.283	0.404	0.641

geomagnetic event if a CME previously impacted Earth. In other words, the closer the forecast future, the simpler the ML-based modeling of the cause-and-effect relationship between any solar driver and geomagnetic activity. The recurrent LSTM architecture performs slightly better than the others (in bold in Table 3), achieving a balanced accuracy of 93.6%, and a precision, sensitivity, and F1 score of 62.0%, 87.6%, and 73.6%, respectively. In contrast, MLP and stacked (deep CNN) networks are the least efficient, suggesting that simply adding multiple connected layers does not give benefits unless solutions aimed at emulating recurrent models are adopted, as in the case of the deep CNN_{skip} architecture. Interestingly, the models' sensitivity (82.8% on average) is generally higher than their precision (63.3% on average) in predicting critical events. This result is consistent with the different importance given to the two classes of events: getting a false alarm (lower precision) may be less harmful than missing a criticality (higher sensitivity).

When considering an 8 hr time horizon for forecasting (MLP and deep CNN architectures were discarded, given their poor results even in the nowcasting test), model performance decreases significantly, although it remains satisfactory (balanced accuracy, precision, sensitivity, and F1 score of 69.5%, 69.4%, 39.1%, and 50.0%, respectively) for the hybrid LSTM-FCN architecture (in bold in Table 4), evidently made more efficient by its recurrent module. This is in line with expectations and has a physics-based explanation. When the CME impacts Earth, it indeed takes some time to cause the geomagnetic disturbance. After the main phase of the solar storm, a more or less slow recovery phase (modulated by the presence of Alfvénic streams following the solar driver; Telloni et al. 2021) restores the Earth's magnetosphere to equilibrium conditions. Therefore, it is clear that predictive models cannot push their forecast time horizon too far into the recovery phase (when the SYM-H is returning above −50 nT) or even into the pre-storm equilibrium state, as there is an inherent constraint related to the timescales of solar-wind–Earth interaction. For the same reason, the precision of the NNs is in this case higher than their sensitivity. Moving the prediction time horizon forward, plausibly as early as the recovery phase, makes the NN more likely to fail to predict a geomagnetic event than to provide a false alarm.

This motivates us to perform a more complicated multistep test trying to simultaneously predict critical events from 1–8 hr ahead, thus allowing the temporal extension of the CME impact on the Earth's magnetosphere to be investigated. The BA for the five best-performing NNs in the previous application is displayed in Figure 2 as a function of the alert timings.

As expected, for all models the accuracy decreases with time, which suggests that different phases (characterized by different SYM-H values) of the geomagnetic storm are involved in the time

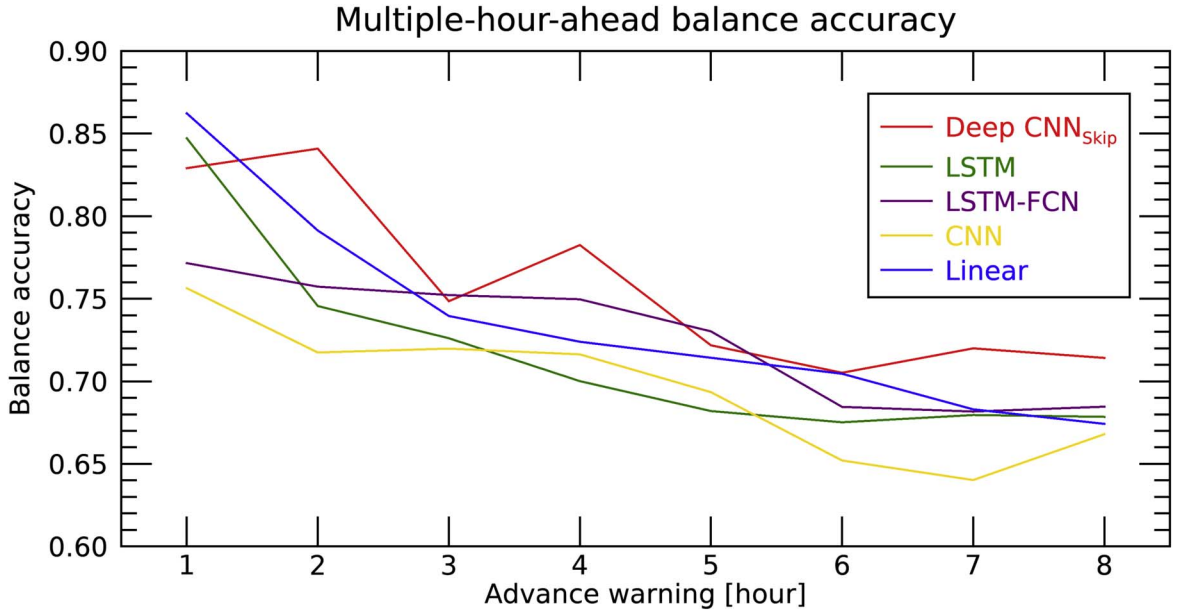

Figure 2. Balance accuracy for the five most efficient NN architectures (marked with different colored lines as reported in the legend) for multiple-hour advance alerts for critical (SYM-H < -50 nT) events.

horizon being predicted. The deep CNN with the skip connection architecture offers the highest performance, ensuring a prediction accuracy greater than 70% for all eight future time instants. What is most interesting, however, is that after an initial decrease, its accuracy peaks at 4 hr. This behavior is also suggested by the improved performance of LSTM-FCN and CNN models in the middle of the considered forecast time range. Although further investigations should be performed to corroborate this result, it can be easily interpreted in relation to the time required by a CME to perturb the geospace. As a matter of fact, Telloni et al. (2019) showed that the average waiting time between the CME detection and the onset of the geomagnetic storm is about 4 hr. From an operational point of view, therefore, it is expected that ML-based algorithms, in addition to being increasingly efficient in the nearest future, will also offer good predictive capabilities at these early warning times (probably enabled by increasing the complexity of neural architectures, as for the deep $\mathrm{CNN_{skip}}$ model). Besides supporting the results of Telloni et al. (2019) and providing insights into the application limitations of space weather nowcasting, this finding also sheds light on the physics behind the generation of ring currents and the corresponding timescales, opening up wide opportunities for the investigation of these as-yet underexplored topics.

As a final remark, it is worth noting that, by testing the performance of all the different types of NNs, the results shown in this section confirm the general practice in the literature of considering (and thus primarily resorting to) recurrent models as those best suited to provide predictions when applied to time series. Further investigations are needed both from a physical perspective to disclose the characteristic timescales of the Sun–Earth interactions and from an application standpoint to provide better predictions of geomagnetic disturbances. In the latter case, the combined use of in situ measurements at L1 and remote coronal observations of CMEs could be beneficial in integrating ML techniques that could jointly forecast the arrival of the CME to Earth (Guastavino et al. 2023) and its geoeffectiveness, thereby improving the reliability and predictive capability of the adopted NNs. In addition, it is imperative to also extend the analysis to include forecasting the severity of geomagnetic storms (i.e., moderate, intense, and extreme, e.g., Cander & Mihajlovic 1998) by adopting a category classification approach. However, this is all devoted to future works.

Acknowledgments

D.T. was partially supported by the Italian Space Agency (ASI) under contract 2018-30-HH.0. S.G. acknowledges the financial support of the Programma Operativo Nazionale (PON) "Ricerca e Innovazione" 2014-2020. V.C. acknowledges the financial support of INdAM-GNCS. The combined solar wind and geomagnetic data analyzed in this paper are public and can be freely downloaded from the NASA's Space Physics Data Facility (http://omniweb.gsfc.nasa.gov/).

ORCID iDs

Daniele Telloni https://orcid.org/0000-0002-6710-8142
Maurizio Lo Schiavo https://orcid.org/0009-0007-0794-1873
Enrico Magli https://orcid.org/0000-0002-0901-0251
Silvano Fineschi https://orcid.org/0000-0002-2789-816X
Sabrina Guastavino https://orcid.org/0000-0001-7047-1148
Gianalfredo Nicolini https://orcid.org/0000-0002-9459-3841
Roberto Susino https://orcid.org/0000-0002-1017-7163
Silvio Giordano https://orcid.org/0000-0002-3468-8566
Francesco Amadori https://orcid.org/0000-0003-1316-1033
Valentina Candiani https://orcid.org/0000-0003-0669-8019
Anna Maria Massone https://orcid.org/0000-0003-4966-8864
Michele Piana https://orcid.org/0000-0003-1700-991X

References

Aggarwal, C. C. 2018, Neural Networks and Deep Learning (Berlin: Springer)
Bala, R., & Reiff, P. 2012, SpWea, 10, S06001
Bala, R., Reiff, P., & Russell, C. T. 2015, JGRA, 120, 3432

Bishop, C. M. 1995, Neural Networks for Pattern Recognition (Oxford: Oxford Univ. Press)
Bobra, M. G., & Ilonidis, S. 2016, ApJ, 821, 127
Burlaga, L. F., & Ogilvie, K. W. 1970, ApJ, 159, 659
Camporeale, E. 2019, SpWea, 17, 1166
Camporeale, E., Carè, A., & Borovsky, J. E. 2017, JGRA, 122, 10,910
Cander, L. R., & Mihajlovic, S. J. 1998, JGR, 103, 391
Collado-Villaverde, A., Muñoz, P., & Cid, C. 2021, SpWea, 19, e02748
Goodfellow, I., Bengio, Y., & Courville, A. 1996, Deep Learning (Cambridge, MA: MIT Press)
Gosling, J. T., McComas, D. J., Phillips, J. L., & Bame, S. J. 1991, JGR, 96, 7831
Gosling, J. T., & Pizzo, V. J. 1999, SSRv, 89, 21
Gruet, M. A., Chandorkar, M., Sicard, A., & Camporeale, E. 2018, SpWea, 16, 1882
Guastavino, S., Candiani, V., Marchetti, F., et al. 2023, ApJ, submitted
He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proc. of 2016 IEEE Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE), 770
Hochreiter, S., & Schmidhuber, J. 1997, Neural Comput., 9, 1735
Hornik, K., Stinchcombe, M., & White, H. 1989, NN, 2, 359
Iwana, B. K., & Uchida, S. 2021, PLoSO, 16, e0254841
Iyemori, T. 1990, JGG, 42, 1249
Karim, F., Majumdar, S., Darabi, H., & Chen, S. 1998, IEEE Access, 6, 1662
Kataoka, R., & Nakano, S. 2021, GeoRL, 48, e96275
Keesee, A. M., Pinto, V., Coughlan, M., et al. 2020, FrASS, 7, 72
King, J. H., & Papitashvili, N. E. 2005, JGRA, 110, A02104
Kitajima, R., Nowada, M., & Kamimura, R. 2022, EP&S, 74, 145
Klein, L. W., & Burlaga, L. F. 1982, JGR, 87, 613
Laperre, B., Amaya, J., & Lapenta, G. 2020, FrASS, 7, 39
Lazzús, J. A., Vega, P., Rojas, P., & Salfate, I. 2017, SpWea, 15, 1068
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
Li, H., Wang, C., Tu, C., & Xu, F. 2020, E&SS, 7, e00997
Li, Y. Y., Huang, S. Y., Xu, S. B., et al. 2022, ApJS, 260, 6
Liu, H., Liu, C., Wang, J. T. L., & Wang, H. 2020, ApJ, 890, 12
Liu, J., Ye, Y., Shen, C., Wang, Y., & Erdélyi, R. 2018, ApJ, 855, 109
Lopez, R. E., & Freeman, J. W. 1986, JGR, 91, 1701
Marchetti, F., Guastavino, S., Piana, M., & Campi, C. 2022, PatRe, 132, 108913
Matthaeus, W. H., Elliott, H. A., & McComas, D. J. 2006, JGRA, 111, A10103
Matthaeus, W. H., & Goldstein, M. L. 1982, JGR, 87, 6011
Nguyen, G., Aunai, N., Fontaine, D., et al. 2019, ApJ, 874, 145
Park, W., Lee, J., Kim, K.-C., et al. 2021, JSWSC, 11, 38
Raju, H., & Das, S. 2021, SoPh, 296, 134
Reiss, M. A., Möstl, C., Bailey, R. L., et al. 2021, SpWea, 19, e2021SW002859
Richardson, I. G., Webb, D. F., Zhang, J., et al. 2006, JGRA, 111, A07S09
Roberts, D. A., Karimabadi, H., Sipes, T., Ko, Y.-K., & Lepri, S. 2020, ApJ, 889, 153
Sabbatini, F., & Grimani, C. 2023, arXiv:2302.06740
Singh, P. K. 2022, InJPh, 96, 2235
Sudar, D., Vršnak, B., & Dumbović, M. 2016, MNRAS, 456, 1542
Tasistro-Hart, A., Grayver, A., & Kuvshinov, A. 2021, JGRA, 126, e28228
Telloni, D., Antonucci, E., Bemporad, A., et al. 2019, ApJ, 885, 120
Telloni, D., Bruno, R., D'Amicis, R., Pietropaolo, E., & Carbone, V. 2012, ApJ, 751, 19
Telloni, D., Carbone, F., Antonucci, E., et al. 2020, ApJ, 896, 149
Telloni, D., D'Amicis, R., Bruno, R., et al. 2021, ApJ, 916, 64
Upendran, V., Cheung, M. C. M., Hanasoge, S., & Krishnamurthi, G. 2020, SpWea, 18, e02478
Webb, D. F., & Howard, T. A. 2012, LRSP, 9, 3
Wintoft, P., & Wik, M. 2021, FrASS, 8, 72
Wu, J.-G., & Lundstedt, H. 1997, JGR, 102, 14255
Yang, Y., & Shen, F. 2021, Univ, 7, 371