

The MIDA's activities in space sciences: Artificial intelligence for space weather

University of Genova Department of Mathematics MIDA group

Sabrina Guastavino

AI in Astronomy Workshop



Catania, Italy May 21th 2025







Who is MIDA?

MIDA - Methods for Image and Data Analysis <u>https://mida.unige.it/</u>



Computational astrophysics Solar physics Space Weather



Computational medicine System biology Medical imaging



Mathematics Computional methods Artificial intelligence



Technology Transfer Industrial Mathematics

MIDA in the space: solar image reconstruction

250

200 ្ឋ

100

250

200

ື້ອ 150

100

250

200

j 150

100

650

ĩ

4-8keV

700 750 X (arcsec)

800

7

j 150



Massa P., et al., The STIX *imaging concept*, Solar Physics 298.10 (2023): 114.

Hard X-Ray imaging



650 700 750 X (arcsec) 800 18-50keV

0

800



Desaturation EUV image reconstruction



<u>Guastavino, S.</u> et al., Desaturating SDO/AIA observations of solar flaring storms, ApJ 882.2 (2019): 109.

Sabrina Guastavino | MIDA Group

Al in Astronomy Workshop, Catania, 21th May 2025

MIDA in the space: space weather



Funded projects:

- FLARECAST "Flare Likelihood And Region Eruption foreCASTing" H2020 project
- AI-FLARE "Artificial Intelligence for the analysis of solar FLARE data" project funded by ASI-INAF
- **SWESNET** Space Weather European Network
- Alxtreme "Physics-based AI for predicting extreme weather and space weather events" project funded by Fondazione Compagnia San Paolo e Fondazione Cassa Depositi e Prestiti
- ARCAFF "Active Region Classification And Flare Forecasting" HORIZON project
- **CAESAR** "Comprehensive spAce wEather Studies for the ASPIS prototype Realization" project funded by ASI-INAF
- **CORNERSTON** Project PRIN 2022 PNRR
- Project "Investigating coronal holes' shapes and their relationship with high-speed solar wind streams via topological data analysis" funded by INAF

From Sun to Earth...what are we interested in forecasting to?

High-energy events, such as **solar flares**, which can generate **Coronal Mass Ejectios (CMEs)** and thus fast solar wind and hence **geomagnetic storms**







From Sun to Earth..

1) Solar flare forecasting



Classification problem

Imbalanced dataset!

- 20% of above C class events
- 5% of above M class events
- 0.1% of above X class events

2) Predicting CMEs' travel time from Sun to Earth



Regression problem

3) Predicting severe geomagnetic storms



Classification problem Imbalanced dataset! 2.5% of positive events (SYM-H < -50nT)

Solar flare forecasting

Solar flares originate from magnetically active regions (ARs) but not all solar ARs give rise to a flare.



Two approaches:

- features, representing physical parameters, can be previously extracted from HMI images and then some machine learning techniques are used for the prediction (as SVM, RF, Hybrid Lasso..)
- features can be automatically extracted by some deep learning methods as Convolutional Neural Networks (CNNs).

Some comments

Feature-based ML [1,2,3,4]

• Effort in a priori feature extraction 🕑

• from a physical viewpoint we can analyse which are the most relevant physical features () Image-based DL [5,6]:

- No a priori feature extraction
- The prediction model is less interpretable from physical point of view (\cdot)

[1] Benvenuto F., et al. A Hybrid Supervised/Unsupervised Machine Learning Approach to Solar Flare Prediction, ApJ, vol. 853, iss. 1, 2018 [2] Campi et al., Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence, ApJ, vol. 883, iss. 2, p. 150, 2019

[3] Benvenuto F., et al., Machine learning as a flaring storm warning machine: was a warning machine for the 2017 September solar flaring storm possible?, ApJ, vol. 904, iss. 1, p. L7, 2020

[4] Cicogna D., et al., Flare-forecasting algorithms based on high-gradient polarity inversion lines in active regions, ApJ, vol. 915, iss. 1, p. 38, 2021

[5]Guastavino et al, *Implementation paradigm for supervised flare forecasting studies: a deep learning application with video data*, 2022 A&A, vol. 662, iss. A105 (2022)

[6] Guastavino S., et al., Operational solar flare forecasting via video-based deep learning, Frontiers in Astronomy and Space Sciences, Vol 9, (2023)

Crucial points in machine-deep learning

Given

- a supervised machine or deep learning method for classification (e.g. a neural network)
- a historical data set to be divided in training, validation and test set

How to split training, validation and test sets?

The chronological splitting introduces a bias due to the cyclicity of the solar cycle.



We formulate a data generation process [1] which is based on machine learning theory: **training, validation and test** sets should be drawn from the same distribution (Vapnik, 1998).

[1] Guastavino et al, Implementation paradigm for supervised flare forecasting studies: a deep learning application with video data, (2022) A&A, vol. 662, iss. A105.

Crucial points in machine-deep learning

Given

- a supervised machine or deep learning method for classification (e.g. a neural network)
- a historical data set to be divided in training, validation and test set

How to train the neural network?

A neural network is a parametric function that approximates the map connecting data to the event probability

$$f(w,\cdot):X\to [0,1]$$

Given the training set $\{(X, Y)\}=\{(x_i, y_i)\}_{i=1}^n$, training the neural networks means solving a minimization problem as follows



Crucial points in machine-deep learning

Given

- a supervised machine or deep learning method for classification (e.g. a neural network)
- a historical data set to be divided in training, validation and test set

How to evaluate performances?

Evaluation metrics must take into account the nature of the problem:

1. Imbalance of dataset → Skill scores as True Skill Statistic (TSS), Heidke Skill Score (HSS)..

$$CM(y,\hat{y}) = \begin{pmatrix} TN(y,\hat{y}) & FP(y,\hat{y}) \\ FN(y,\hat{y}) & TP(y,\hat{y}) \end{pmatrix} \quad TSS(CM(y,\hat{y})) = \frac{TP(y,\hat{y})}{TP(y,\hat{y}) + FN(y,\hat{y})} + \frac{TN(y,\hat{y})}{TN(y,\hat{y}) + FP(y,\hat{y})} - 1$$

2. The temporal aspect of predictions \rightarrow Value-weighted skill scores [1]

[1] Guastavino, **Piana**, **Benvenuto**, Bad and Good Errors: Value-Weighted Skill Scores in Deep Ensemble Learning, (2022) IEEE Transactions on Neural Networks and Learning Systems

Solar flare forecasting: Deep neural network

Input: videos of HMI magnetograms

Deep learning technique:

Long-term Recurrent Neural Network LRCN = CNN + LSTM

Guastavino et al, Implementation paradigm for supervised flare forecasting studies: a deep learning application with video data, (2022) A&A, vol. 662, iss. A105.



Gap between loss minimization and score maximization

In binary classification a common choice of the loss function is the binary crossentropy (CE) $L(f(\mathbf{w}, \mathbf{X}), \mathbf{Y}) = -\sum_{i=1}^{n} y_i \log(f(\mathbf{w}, x_i)) + (1 - y_i) \log(1 - f(\mathbf{w}, x_i))$



[1] Marchetti, Guastavino, Piana, Campi, Score-Oriented Loss (SOL) functions (2022), Pattern Recognition, Vol. 132

Sabrina Guastavino | MIDA Group

Ingredients

SDO/HMI images recorded in the time range between 2012 September 14 and 2017 September 30

- For each AR, we consider the HMI magnetogram frames associated to it and we organize them in 24 hour long time series of HMI magnetogram frames.
- we have a collection of samples, each one represented by a **video** of HMI magnetogram frames associated to an AR.
- data from the past: we know if a flare occurred w.r.t. each time series, therefore we can label each time series with 0 if no flare occurred and with 1 if flares occurred.
- We generate 10 training, validation and test accordingly to a uniform-based splitting strategy in order to assess the statistical robustness in results

Results

				TSS $(C1+$	flares)		
	Mean	Std	Min	25th perc	Median	75th perc	Max
Validation	0.57	0.02	0.55	0.56	0.57	0.59	0.61
Test	0.55	0.05	0.46	0.52	0.54	0.60	0.61
		TSS $(M1 + flares)$					
	Mean	Std	Min	25th perc	Median	75th perc	Max
Validation	0.76	0.07	0.65	0.67	0.77	0.82	0.85
Test	0.68	0.09	0.55	0.61	0.69	0.72	0.82

Coronal mass ejections' travel time prediction

Coronal Mass Ejections (CMEs) consist of large eruptions of plasma and magnetic field that are typically triggered by solar flares and they propagate from the solar corona into the heliosphere.

The observations of CMEs are typically performed by means of remote-sensing instruments

Large Angle and Spectrometric Coronagraph (**LASCO**) on board the Solar and Heliophysics Observatory (SOHO)) Once the CME is detected from LASCO images

When will the CMF

reach the Earth?

Feature extraction

In 48 hours

CMEs' kinematic parameters (as the initial propagation speed, the CME mass, and the initial cross section)

How to face the problem?

- Model-driven approaches (drag-based model, MHD-based equations..)
- Data-driven approaches (artificial intelligence: machine or deep learning)

Physics-driven machine learning

 Model-driven approaches (drag-based model, MHD-based equations..) Data-driven approaches

 (artificial intelligence: machine or deep learning)

How to combine them?

Physics-driven machine learning [1]

Starting from the physics...

Drag-based model

~ 12 h

Typical mean absolute

error in the prediction

$$\ddot{r}(t) = -\gamma |\dot{r}(t) - w| (\dot{r}(t) - w)$$

Drag parameter $\gamma = C \frac{A\rho}{m}$

ρ = solar wind density
A = CME impact area
m = CME mass
C = drag coefficient (unknown)

 \ddot{r} = CME acceleration \dot{r} = CME speed w = solar wind speed

Drag Equation is completed to a Cauchy problem by including the two initial conditions $r(t_0) = r_0$ $\dot{r}(t_0) = v_0$

where r_0 is the height of the eruption ballistic propagation, and v_0 is the initial CME speed.

[1] **Guastavino et al.,** *Physics-driven machine learning for the prediction of coronal mass ejections' travel times,* ApJ, 954:151 (9pp), 2023

Drag-based model

In fact, assuming that the solar wind speed and the drag parameter are constant and homogeneous, the drag
equation leads to

$$\dot{r}(t) = \frac{v_0 - w}{1 + \gamma \operatorname{sign}(v_0 - w)(v_0 - w)t} + w$$
$$r(t) = \operatorname{sign}(v_0 - w)\frac{1}{\frac{A}{m}C\rho} \log\left(1 + \frac{A}{m}C\rho\operatorname{sign}(v_0 - w)(v_0 - w)t\right) + wt + r_0$$

This equation can be used to estimate the travel time as the solution of r(t) = 1 AU.

this approach is reliable just if accurate estimates of the parameters A, m, C, ρ , w, r_0 and v_0 are at disposal.

Specifically, measurements of A, m, r_0 , and v_0 can be provided by coronagraphic instruments such as LASCO. ρ and w at the CME onset can be provided by in situ instruments (such as WIND, ACE, and CELIAS). These same values can be used in the equations as approximations of average values of the two parameters

Physics-driven machine learning

Idea:
Include the drag-based model in the loss function!
$$sign(v_0 - w) \approx \frac{(v_0 - w)}{\sqrt{(v_0 - w)^2 + \delta}}$$
 $x = (v_0, m, A, \rho, \omega)$ $r(t, C) = \frac{1}{\frac{A}{m}C\rho\sigma}\log\left(1 + \frac{A}{m}C\rho\sigma(v_0 - w)t\right) + wt + r_0$ $t = travel time$ λ is a regularization parameterwhere $\sigma = \frac{(v_0 - w)}{\sqrt{(v_0 - w)^2 + \delta}}$ $L_t(t, f(w, x)) = \lambda \frac{(t - f(w, x))^2}{Data-driven term} + (1 - \lambda) \frac{(1 - r(f(w, x), C))^2}{Physics-driven term}$

 $\lambda = 1 \rightarrow \text{only data-driven term} \rightarrow \text{Fully data-driven}$ $\lambda = 0 \rightarrow \text{only physics-driven term} \rightarrow \text{Fully physics-driven}$ $\lambda \in (0,1) \text{ (e.g. } \lambda = 0.5) \rightarrow \text{ both terms} \rightarrow \text{Mix}$

An open issue: how to estimate C?

Physics-driven neural networks



Physics-driven neural networks



Dataset

•	123 CME events	occurred in th	ne time range	between 199	7 and 2018	[1]
---	----------------	----------------	---------------	-------------	------------	-----

Name	Notation	Unity	Description	Source
CME height of eruption	r_0	km	$r_0 = 20 \ R_{\odot}, \ R_{\odot} = 6.957 \cdot 10^5 \ {\rm km}$	-
CME time of eruption	t_0	s	eruption time on the Sun at r_0	(Napoletano et al. 2022)
CME Time of Arrival	ToA	s	estimated arrival time at 1 AU	R & C
CME Travel time	TT	s	estimated time between t_0 and ToA	R & C,
				(Napoletano et al. 2022)
CME initial speed	v_0	$\rm km/s$	initial propagation speed from eruption	LASCO
CME mass	m	g	estimated CME mass	LASCO
CME impact area	A	km^2	CME impact area, constant angular width	LASCO
Solar wind density	ho	$ m g/km^3$	mean over one hour after t_0	CELIAS
Solar wind speed	w	m km/s	mean over one hour after t_0	CELIAS
Drag parameter	C	$\operatorname{dimensionless}$	parameter of the drag based model	this work

 In order to perform a statistical assessment of the physics-driven machine learning approach to travel time prediction, we realized 100 random realizations of the training, validation, and test

[1] Napoletano, G, et al. "Parameter Distributions for the Drag-Based Modeling of CME Propagation." Space Weather 20.9 (2022): e2021SW002925

Data-driven versus physics-driven machine learning

Fully data-driven N1 is switched off C is not an input of the second N2 $\lambda = 1$

$$L_T(t, N2(x)) = (t - N2(x))^2$$

Physics-driven approach N1 is switched on C is an input of the second N2 $\lambda = 0.5$ $L_T(t, N2(\bar{x})) = \frac{1}{2} (t - N2(\bar{x}))^2 + \frac{1}{2} (1 - r(N1(x), N2(\bar{x})))^2$





Geomagnetic storm prediction

Aim: predict the occurrence of severe geomagnetic event (**SYM-H < - 50nT**) in the **next hour** starting **from 24-hour long time series of in-situ measurement** with cadence 1 hour

Dataset

<u>Source</u>: ACE and WIND <u>Temporal range</u>: 15 years (from 2005 to 2019) cadence 1 min #samples = 7 888 320

From in situ measurement:

- B, Bx, By, Bz,
- V, Vx, Vy, Vz
- Density

Label:

SYM-H

- Temperature
- KineticEnergy, MagneticEnergy, TotalEnergy,
- MagneticHelicity

The first indication of a geomagnetic storm is shown by a decrease of the SYM-H index



storms are classified as small if -50 nT < SYM-H ≤ -30 nT moderate if -50 nT < SYM-H < -100 nT intense if SYM-H ≤ -100 nT



Derivative

quantities



Geo-effectiveness prediction as a binary classification problem



[1] Telloni, D., et al, *Prediction Capability of Geomagnetic Events from Solar Wind Data Using Neural Networks*, ApJ, 952:111 (8pp), 2023

[2] Marchetti, F., Guastavino, S., et al, Score-oriented loss functions, ApJ, 952:111 (8pp), 2023

Feature ranking

Different feature ranking methods were explored:

- Correlation-driven permutation importance method [1]
- Greedy-feature selection method [2]

Correlation-driven feature selection method [1] to rank the most relevant features involved in the neural network prediction model. Such method is based on two principles:

- evaluating how each feature impacts the predictions according to some permutation importance score: how the accuracy of the prediction changes when a single feature is randomly shuffled in the validation dataset
- 2. including the absolute correlation coefficients between features in the permutation process.

[1] Kaneko, H., Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances*, *3*(9-10), 278-287 (2022).

[2] Camattari*, **Guastavino***, et al, Classifier-dependent feature selection via greedy methods, Statistics and Computing 34:151 (2024).

[3] **Guastavino S.,** et al., Forecasting Geoffective Events from Solar Wind Data and Evaluating the Most Predictive Features through Machine Learning Approaches, ApJ 971 94 (2024)

Ranking [3]

Feature	Importance value		
1) SYM-H	0.5577		
2) B	0.3392		
3) Total Energy	0.3334		
4) Vx	0.3256		
5) V	0.3162		
6) Temperature	0.287		
7) Magnetic Energy	0.2793		
8) Magnetic Helicity	0.2628		
9) Kinetic Energy	0.2427		
10) Bz	0.2303		
11) Density	0.2044		
12) Vy	< 0.1		
13) By	< 0.1		
14) Vz	< 0.1		
15) Bx	< 0.1		

All features vs selected features



Guastavino S., et al., Forecasting Geoffective Events from Solar Wind Data and Evaluating the Most Predictive Features through Machine Learning Approaches, (2024) ApJ

May Superstorm 2024

In 2024 May 10/11, the second strongest geomagnetic storm has occurred in the space era, with a peak Dst index of lower than -400 nT. The storm was caused by NOAA Active Region (AR) 13664, which was the source of a large number of coronal mass ejections and flares, including more than 10 X-class flares.



May Superstorm 2024



Solar storm forecasting



CME's travel time prediction



In-situ data and geomagnetic storm

Input: 24-hour time series (sampled at 1 hour) of the following 10 selected features

- B
- Bz
- V
- Vx
- Temperature
- KineticEnergy
- MagneticEnergy
- TotalEnergy,
- MagneticHelicity
- SYM-H

AI method: LSTM with TSS-score oriented loss function

Guastavino S., et al., Forecasting Geoffective Events from Solar Wind Data and Evaluating the Most Predictive Features through Machine Learning Approaches, (2024) ApJ



Geomagnetic storm prediction



Summary and conclusions

Flare forecasting

- Benvenuto F., Piana M., Campi C., Massone A. M., A Hybrid Supervised/Unsupervised Machine Learning Approach to Solar Flare Prediction, ApJ, vol. 853, iss. 1, 2018
- Campi et al., Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence, ApJ, vol. 883, iss. 2, p. 150, 2019
- Piana M., Campi C., Benvenuto F., <u>Guastavino S.</u>, Massone A. M., *Flare forecasting and feature ranking using SDO/HMI data*, Nuovo Cimento della Società Italiana di Fisica c, vol. 42, iss. 1, 2019
- Benvenuto F., Campi C., Massone A. M., Piana M., *Machine learning as a flaring storm warning machine: was a warning machine for the 2017 September solar flaring storm possible?*, ApJ, vol. 904, iss. 1, p. L7, 2020
- Cicogna D., Berrilli F., Calchetti D., Moro D. D., Giovannelli L., Benvenuto F., Campi C., <u>Guastavino S.</u>, Piana M., *Flare*forecasting algorithms based on high-gradient polarity inversion lines in active regions, ApJ, vol. 915, iss. 1, p. 38, 2021
- <u>Guastavino S.</u>, Marchetti F., Benvenuto F., Campi C., Piana M., *Implementation paradigm for supervised flare forecasting studies: a deep learning application with video data*, A&A, vol. 662, iss. A105, 2022
- <u>Guastavino, S.</u>, Piana, M., and Benvenuto, F., *Bad and Good Errors: Value-Weighted Skill Scores in Deep Ensemble Learning*, IEEE Transactions on Neural Networks and Learning Systems, pp.1-10, 2022
- <u>Guastavino S.</u>, Marchetti F., Benvenuto F., Campi C., Piana M., *Operational solar flare forecasting via video-based deep learning*, Frontiers in Astronomy and Space Sciences, Vol 9, 2023

Summary and conclusions

Coronal Mass Ejections' travel time prediction

- <u>Guastavino, S.</u> et al., *Physics-driven machine learning for the prediction of coronal mass ejections' travel times*, ApJ, 954:151 (9pp), (2023)
- Rossi, M., <u>Guastavino, S.</u>, Piana, m., Massone, A.M., Extended-drag based model, A&A

Geomagnetic storm prediction

- Telloni D., Lo Schiavo M., Magli E., Fineschi S., <u>Guastavino S.</u>, Nicolini G., Susino R., Giordano S., Amadori F., Candiani V., Massone A. M., Piana M., Prediction Capability of Geomagnetic Events from Solar Wind Data Using Neural Networks, Apj, (2023) Vol. 952, Number 2
- <u>Guastavino S.</u>, et al., Forecasting Geoffective Events from Solar Wind Data and Evaluating the Most Predictive Features through Machine Learning Approaches, ApJ 971 94 (2024)
- Camattari F.*, <u>Guastavino S.*</u>, Marchetti F., Piana M., Perracchione E., Classifier-dependent feature selection via greedy methods, Statistics and Computing 34, 151 (2024)

Summary and conclusions

Active region classification and localization

Legnaro, E., <u>Guastavino, S.</u>, Piana, M., Massone, A.M., *Deep Learning for Active Region Classification: A Systematic Study from Convolutional Neural Networks to Vision Transformers*, ApJ **981** 157 (2025)

May Superstorm prediction

- <u>Guastavino S.,</u> Legnaro E., Massone A.M., Piana M., Artificial Intelligence for the Characterization of the May 2024 Superstorm:Active Region Classification, Flare Forecasting, and Geomagnetic Storm Prediction, ApJ (2025)
- <u>Guastavino S.</u>, Legnaro E., Massone A.M., Piana M., Physics-driven machine learning can estimate the travel time of one of the May 2024 Coronal Mass Ejections with surprisingly high accuracy submitted to ApJL





Thank you!



e-mail: sabrina.guastavino@unige.it

MIDA Group | University of Genova, Italy







Fondazione Compagnia di SanPaolo