Exploring the Universe with Artificial Intelligence (UniversAI)
Proceedings IAU Symposium No. 397, 2026
I. Liodakis & V. Efthymiou, eds.
doi:10.1017/S1743921325104699

SWAAT: An Automated Framework for Space Weather Data Analysis and Prediction

Felipe Meza-Obando[1,3,4], Juan Luis Crespo-Mariño[1,4], Carolina Salas-Matamorros[2,3] and Nicolás Ruiz-Carrazco[4]

[1]Escuela de Ingeniería Mecatrónica, Tecnológico de Costa Rica
email: fmeza@itcr.ac.cr

[2]Space Research Center, Universidad de Costa Rica

[3]Laboratorio de Astrofísica Solar y Clima Espacial, Universidad de Costa Rica

[4]Laboratorio de Inteligencia Artificial para las Ciencias Naturales, Tecnológico de Costa Rica

Abstract. Predicting interplanetary coronal mass ejections (ICMEs) remains challenging due to limited datasets, the complexity of some approaches based on machine learning and the continued reliance on manual event validation. We introduce SWAAT (Space Weather Automated Analysis Tool), a framework that integrates automated validation with synthetic data-driven prediction to improve ICME transit time (TT) forecasts based on propagation velocity (Vp). The system comprises two main modules: (i) The Automated Data Acquisition and Analysis (ADAA), which automates ICME candidate validation using GOES X-ray and SDO imagery, and (ii) The Generative-Evolutionary Synthetic Data (GESD), which employs evolutionary algorithms to generate optimized synthetic datasets for model training. By combining automatic event validation with quality synthetic data, SWAAT mitigates the limitations imposed by scarce datasets thanks to the contribution of synthetic data, improving the accuracy of space weather forecasting.

Keywords. Space Weather, ICME Prediction, Machine Learning, Automated Data Analysis, Generative Evolutionary, Synthetic Data, Solar Analysis, Coronal Mass Ejections.

1. Introduction

ICMEs are among the most impactful solar phenomena for near Earth space. Expelled from the solar corona at speeds often between 20 and 2500 km/s Webb & Howard (2012), they can reach Earth in one to two days, driving severe geomagnetic storms, disrupting communications and navigation systems, and affecting space and ground based infrastructures Schrijver (2015). Accurate estimation of TT is therefore essential for early warning and risk mitigation.

Despite advances in physical and empirical modeling, TT prediction remains challenging. Traditional approaches such as manual catalogs or velocity distance relations are constrained by computational cost, and dependence on expert analysis, while the dynamic interplanetary medium introduces further uncertainties. Machine learning methods have been increasingly applied in space weather over the past few years Lugaz et al. (2021), but their progress remains constrained by the limited availability of datasets with essential variables and by the incompleteness of observational records.

These challenges underscore the need for flexible, data driven frameworks. SWAAT addresses this demand by integrating automated event validation with synthetic data generation guided by an evolutionary algorithm, which mitigates common issues in

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union.

https://doi.org/10.1017/S1743921325104699 Published online by Cambridge University Press

synthetic creation such as overfitting and oversampling, while ensuring low error. Its hybrid architecture combining evolutionary and generative models in the prediction stage with automated processing of GOES/XRS time series and SDO imagery in the validation stage enhances automation and strengthens predictive accuracy.

2. Background

A central challenge in space weather prediction is the scarcity of high quality, reliable data. Although solar missions have produced extensive observations, ICME events remain relatively rare, limiting the massive training of robust machine learning models. Existing catalogs, often based on expert manual validation, ensure quality but slow down the delivery of results.

Traditionally, event analysis has relied on IDL, while in this work we employ Python and multiple scientific libraries, with SunPy Barnes et al. (2023) as the most widely used.

In astrophysics, the use of generative and evolutionary AI remains limited compared to supervised methods already applied in tasks such as galaxy classification, exoplanet detection, and solar structure segmentation. Generative models have been explored mainly for image enhancement or classification support, while evolutionary algorithms have been restricted to parameter tuning rather than dataset generation Zhang et al. (2024).

3. System Overview

The architecture of SWAAT (Fig. 1) is organized into two main processes: the Automatization Process and the Learning Process, each subdivided into functional stages.

• Automatization Process: Candidate events are analyzed by the ADAA module, which validates their significance using solar X-ray and imaging data. Non relevant events are discarded, while validated events trigger both the automatic report generation module and the extraction of the Vp.

• Learning Process: This process integrates two stages. The GESD expands the set of confirmed events by producing synthetic datasets optimized through evolutionary strategies. These enriched datasets feed into the Prediction Stage, where a polynomial regression algorithm estimates the ICME TT.

Together, these processes ensure a consistent workflow: ADAA provides reliable validation, GESD enriches the training base, and the prediction stage delivers the final TT forecast.

3.1. *ADAA: Automation Component*

ADAA is responsible for the early detection and validation of candidate solar events. It operates on two key data sources: GOES X-ray observations and SDO/AIA imagery.

GOES/XRS time series are analyzed in the high energy channel (0.1-0.8 nm), where rapid increases above a predefined threshold (M or X class) are flagged as potential candidates. Noise filtering, peak detection, and temporal derivatives are applied to discriminate genuine events from background fluctuations. In parallel, automated processing of SDO imagery defines a Region of Interest (ROI) covering 90% of the solar disk and identifies Points of Interest (POIs) as localized active regions. The temporal evolution of each POI is monitored, and a discrete derivative is computed to obtain a variability index.

An event is validated only when both criteria are satisfied: (i) the X-ray intensity exceeds the threshold, and (ii) POI variability exceeds the predefined threshold within the same time interval as the XRS peak. Valid events are then documented in automated HTML reports containing temporal plots and segmentation maps, while non events are discarded.

https://doi.org/10.1017/S1743921325104699 Published online by Cambridge University Press

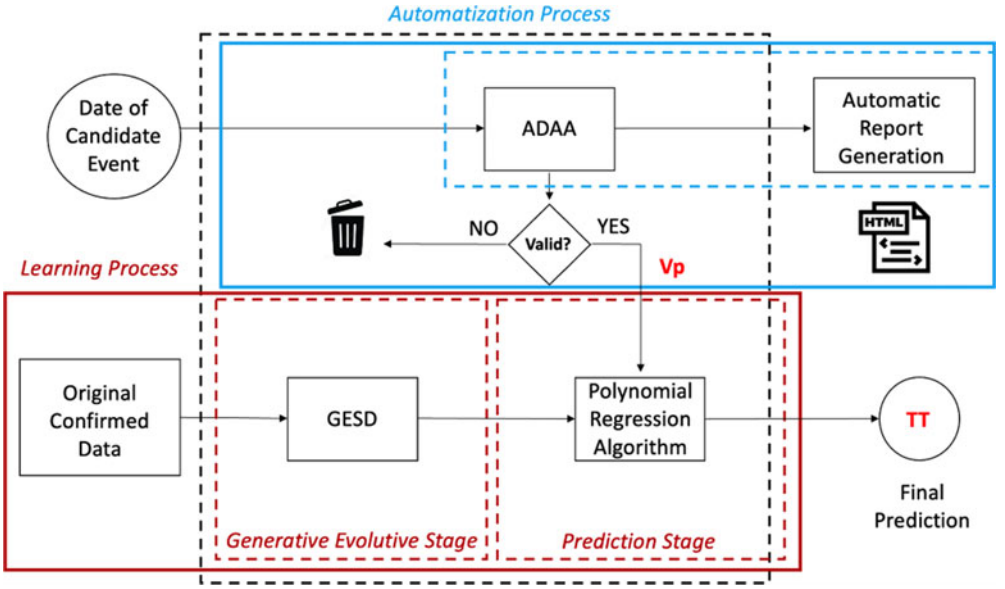


Figure 1. Architecture of SWAAT, showing the Automatization Process for event validation (top) and the Learning Process with synthetic data generation and prediction (bottom).

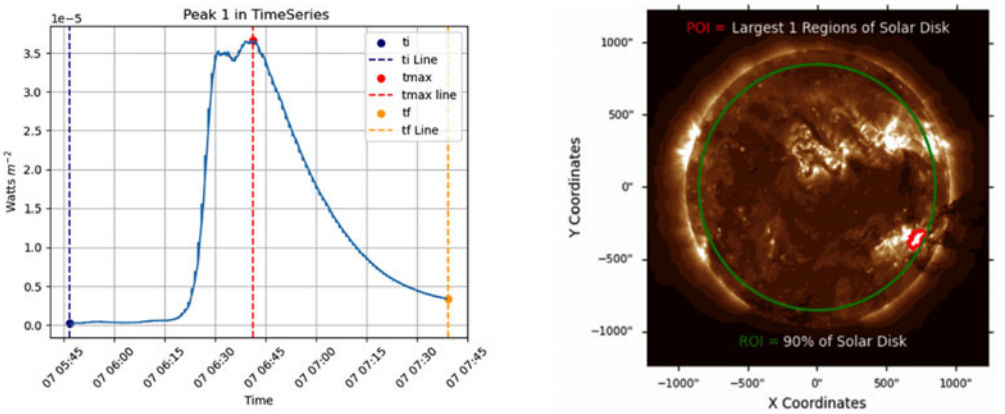


Figure 2. GOES/XRS time series with event initial time (ti) and maximum peak (tmax) (left), and SDO imagery showing the Region of Interest in green (90% of the solar disk) and detected Points of Interest in red (right).

This dual condition process ensures consistent and reproducible validation, providing a reliable entry point to the SWAAT predictive pipeline.

Finally, when an event is validated, the Vp is derived following Salas-Matamoros & Klein (2015), by computing the soft X-ray fluence as the area under the GOES/XRS curve between the initial time (ti) and the maximum pear (tmax).

In Figure 2 the left panel illustrates the identification of solar flare peaks in GOES/XRS time series using a peak detection algorithm that determines the initial time (ti) and maximum peak at tmax. The right panel shows the extraction of Points of Interest (POIs) from SDO imagery through contour based analysis within the Region of Interest (90% of the solar disk). The temporal evolution of POIs is tracked, and a discrete derivative is

https://doi.org/10.1017/S1743921325104699 Published online by Cambridge University Press

applied to these measurements to compute a variability index, which serves as a criterion for event validation.

3.2. *GESD: Synthetic Data Generation through Generative-Evolutionary Methods*

The GESD component is one of the most important elements of SWAAT's learning stage. Here, a high quality synthetic dataset is generated to expand and diversify the original dataset, which are then used in the next step to predict ICME TT through polynomial regression. The synthetic generation process is guided by an evolutionary function that regulates the degree of the polynomial and ensures minimal overfitting, avoids synthetic overpopulation, and drives the model toward the lowest possible error. This capability is essential for overcoming the scarcity of data, a major limitation for machine learning models in predicting ICME TT.

The technical design and validation of GESD have been presented in detail in a previous study Meza-Obando et al. (2024). In this article, we provide only a functional summary and emphasize its integration into the operational workflow of SWAAT. For a full description, the reader is referred to the original publication.

4. Conclusions

The ADAA module has shown promising and consistent results in the automated detection and validation of solar events. It identified $> 95\%$ of significant X-ray peaks in GOES/XRS confirmed by manual inspection and correlated them with localized activity in SDO imagery. Within the defined ROI, the system successfully extracted POIs and quantified their temporal variability using a variability index derived from discrete derivative analysis, reducing validation time from 5-7 hours to < 35 minutes. This provides a robust and reproducible criterion for confirming candidate ICME events, closely matching the behavior expected from expert driven analyses.

A first detailed case study, based on a manually examined ICME event, confirmed that ADAA reproduces key observational features with high fidelity, both in X-ray dynamics and in the evolution of active coronal regions. However, due to the inherent complexity and time consuming nature of full manual analyses, the validation process is still ongoing. Additional events are currently being tested against carefully established manual baselines, ensuring that ADAA's performance is benchmarked rigorously before being fully integrated into operational workflows.

The GESD module, responsible for TT prediction, expanded the training dataset from 86 real events to 260 synthetic+real instances, overcoming the scarcity of observational records. Models trained with this enriched dataset reduced the MAE in ICME TT estimation to 9.7 h, compared with > 10 h using other approaches. This demonstrates the scientific value of SWAAT: by uniting automated validation with informed synthetic data generation, it provides a reliable, reproducible, and operationally relevant framework for improving ICME forecasting.

5. Discussion

A key contribution of SWAAT is mitigating the scarcity of ICME datasets, defined only by TT and Vp, through synthetic data generation, avoiding the need for multiple or complex features. In this framework, generative models are guided by an evolutionary strategy, ensuring that synthetic datasets remain consistent while avoiding common issues in synthetic generation and reducing prediction error. This approach expands the dataset and decreases forecasting error in ICME TT estimation. Beyond space weather, the strategy also provides a flexible methodology for other astrophysical domains facing similar data limitations.

https://doi.org/10.1017/S1743921325104699 Published online by Cambridge University Press

Despite its advances, SWAAT remains limited by its dependence on continuous X-ray data from GOES and solar imagery from SDO. Interruptions, resolution variability, or data gaps can compromise both detection and prediction. Although redundancy was considered at the design stage, the current system lacks multisatellite integration and robust strategies for handling missing data, highlighting clear directions for future improvement.

A promising development for SWAAT is its integration with the Radio Observatorio Santa Cruz (ROSAC), the first solar dedicated radio observatory in Central America. Since ground based radio flux measurements are strongly correlated with X-ray outputs, the same procedure for estimating Vp can be applied. ROSAC will therefore extend SWAAT's observational base and enhance ICME prediction. Technically, this integration will require new pipelines and calibration routines, but it will also reduce dependence on satellites, strengthening the system's capacity for near real time space weather monitoring.

References

Barnes, W.T., Christe, S., et al. 2023, *Front. Astron. Space Sci.*, 10, 1076726.
Lugaz, N., Liu, H., Hapgood, M., & Morley, S. 2021, *Space Weather*, 19.
Meza-Obando, F., Calderón-Quesada, J., et al. 2024, *IEEE CONCAPAN XLII*, 1.
Salas-Matamoros, C., & Klein, K.-L. 2015, *Sol. Phys.*, 290, 1337.
Schrijver, C.J. 2015, *Adv. Space Res.*, 55, 2745.
Webb, D.F., & Howard, T.A. 2012, *Living Rev. Sol. Phys.*, 9, 3.
Zhang, H.L., Wang, J., Zhang, Y.Z., et al. 2024, *Astron. Tech. Instrum.*, 1, 1.

https://doi.org/10.1017/S1743921325104699 Published online by Cambridge University Press