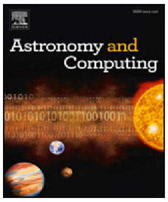
Astronomy and Computing 48 (2024) 100857

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom



Astronomy and Computing

Full length article

Predicting sunspot number from topological features in spectral images I: Machine learning approach

D. Sierra-Porta [a,*], M. Tarazona-Alvarado [b], D.D. Herrera Acevedo [c]

[a] *Universidad Tecnológica de Bolívar, Facultad de Ciencias Básicas, Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena de Indias, 130010, Bolívar, Colombia*
[b] *Universidad Industrial de Santander, Escuela de Física, Car 27 #9, Bucaramanga, 680001, Santander, Colombia*
[c] *Universidad Tecnológica de Bolívar, Facultad de Ingeniería, Parque Industrial y Tecnológico Carlos Vélez Pombo Km 1 Vía Turbaco, Cartagena de Indias, 130010, Bolívar, Colombia*

ARTICLE INFO

Keywords:
Machine learning
Sunspots prediction
Spectral images
Sun's dynamics
Fractal features

ABSTRACT

This study presents an advanced machine learning approach to predict the number of sunspots using a comprehensive dataset derived from solar images provided by the Solar and Heliospheric Observatory (SOHO). The dataset encompasses various spectral bands, capturing the complex dynamics of solar activity and facilitating interdisciplinary analyses with other solar phenomena. We employed five machine learning models: Random Forest Regressor, Gradient Boosting Regressor, Extra Trees Regressor, Ada Boost Regressor, and Hist Gradient Boosting Regressor, to predict sunspot numbers. These models utilized four key heliospheric variables — Proton Density, Temperature, Bulk Flow Speed and Interplanetary Magnetic Field (IMF) — alongside 14 newly introduced topological variables. These topological features were extracted from solar images using different filters, including HMIIGR, HMIMAG, EIT171, EIT195, EIT284, and EIT304. In total, 60 models were constructed, both incorporating and excluding the topological variables. Our analysis reveals that models incorporating the topological variables achieved significantly higher accuracy, with the r2-score improving from approximately 0.30 to 0.93 on average. The Extra Trees Regressor (ET) emerged as the best-performing model, demonstrating superior predictive capabilities across all datasets. These results underscore the potential of combining machine learning models with additional topological features from spectral analysis, offering deeper insights into the complex dynamics of solar activity and enhancing the precision of sunspot number predictions. This approach provides a novel methodology for improving space weather forecasting and contributes to a more comprehensive understanding of solar-terrestrial interactions.

1. Introduction

The investigation into solar activity and its repercussions on Earth's climate (Solomon et al., 2019; Le Mouël et al., 2019; Floyd et al., 2002; Zhang et al., 2021; Singh and Bhargawa, 2020) and technology (Spencer et al., 2019) has been a focal point of extensive research for many years. Sunspots, characterized by regions of reduced brightness on the Sun's surface (Tlatov, 2022; Nandy, 2021), serve as a crucial indicator of solar activity. They are associated with various phenomena such as solar flares, coronal mass ejections, and geomagnetic storms (Gour et al., 2021; Cliver et al., 2022; Alexakis and Mavromichalaki, 2019). Accurately forecasting the quantity of sunspots is, therefore, essential for understanding and mitigating the impacts of solar activity on our planet.

Predicting the number of sunspots is crucial for several reasons. Sunspots are key indicators of solar activity and are instrumental in determining solar cycles, including their maxima, minima, and overall duration. Understanding solar cycles allows scientists to predict space weather events, which can have significant impacts on Earth's climate and technological systems (National Research Council and Division on Engineering and Physical Sciences and Space Studies Board and Committee on the Societal and Economic Impacts of Severe Space Weather Events and A Workshop, 2009; National Research Council and Division on Engineering and Physical Sciences and Aeronautics and Space Engineering Board and Space Studies Board and Committee on a Decadal Strategy for Solar and Space Physics (Heliophysics), 2013). For instance, high solar activity can lead to geomagnetic storms that disrupt communication and navigation systems, damage satellites, and affect power grids. Additionally, long-term variations in solar activity have been linked to climate patterns on Earth (Drews et al., 2022; Tsiropoula, 2003), such as temperature fluctuations and atmospheric

[*] Corresponding author.
E-mail address: dporta@utb.edu.co (D. Sierra-Porta).

https://doi.org/10.1016/j.ascom.2024.100857
Received 12 March 2024; Accepted 15 July 2024
Available online 19 July 2024

circulation changes. Accurate sunspot predictions contribute to better preparedness for these events, thereby mitigating their adverse effects on society and infrastructure.

In astronomy, predicting solar cycles is also crucial for the planning of surveys and nighttime observation schedules to minimize interference from cosmic rays and scattered plasma in the interplanetary medium (Grauer and Grauer, 2021; Barentine, 2022). Accurate predictions enable astronomers to select optimal observation times, reducing noise and enhancing data quality. Additionally, understanding solar cycles aids in developing resilience and adaptive strategies for extreme events that impact terrestrial technological systems, ensuring the continued functionality of communication, navigation, and power infrastructures during periods of intense solar activity.

Recent advancements have shown significant improvement in the use of machine learning models to predict sunspot numbers based on an array of solar parameters (Khan et al., 2020; Xiao, 2021) but also deep learning methods (Pala and Atici, 2019; Prasad et al., 2023; Li et al., 2021) including proton density, temperature, field magnetic average (FMA), and bulk flow speed. Traditionally, these models rely on datasets provided by space missions that measure various characteristics of the photosphere, heliosphere, magnetic fields, and chromosphere. Notable examples include NASA's OMNIWEB services (https://omniweb.gsfc.nasa.gov/ow.html) and spacecraft missions such as NASA's Advanced Composition Explorer (ACE) (Stone et al., 1998; Hill et al., 2020), Wind missions (Wilson et al., 2021), and NOAA's Deep Space Climate Observatory (DSCOVR) mission (Burt and Smith, 2012; Marshak et al., 2018), situated at Lagrange point L1. These missions routinely measure the Interplanetary Magnetic Field (IMF) and other aspects of solar dynamics.

For instance, Dani and Sulistiani (2019) aimed to predict the peak time and value of Solar Cycle 25 using four different machine learning regression methods: Linear Regression (LR), Random Forest (RF), Radial Basis Function (RBF), and Support Vector Machine (SVM). This study utilized monthly mean sunspot number data from 1856 to June 2018 (solar cycles 10–24) provided by the World Data Center SILSO. The RF prediction suggested a lower maximum with a well-defined double-peak, and all methods predicted Solar Cycle 25 to commence in late 2019 or early 2020.

Similarly, Mahdi et al. (2019) explored the relationship between sunspot counts and coronal mass ejections (CMEs) using various classification algorithms, including decision tree, nearest neighbor, support vector machine, discriminant, ensembles, and logistic regression. They found that the Ensemble Bagged Tree model exhibited the best performance with an accuracy of 90.8%.

Moreover, Dang et al. (2022) introduced a novel ensemble model, XGBoost-DL, which integrates deep learning models with XGBoost. This model demonstrated exceptional forecasting performance, surpassing other models with a Root Mean Square Error (RMSE) of 25.70 and Mean Absolute Error (MAE) of 19.82, highlighting the efficacy of ensemble models in sunspot number prediction.

Recently, Tarazona-Alvarado and Sierra-Porta (2023) developed a comprehensive dataset based on solar images from the Solar and Heliospheric Observatory (SOHO). This multidisciplinary effort resulted in a robust methodology for calculating spectral parameters and relevant features from SOHO images, extracting 14 topological features and fractal metrics correlated with sunspot numbers. These metrics include entropy, mean intensity, standard deviation, skewness, kurtosis, relative smoothness, uniformity, fractal dimension, Taruma contrast, Taruma directionality, Taruma coarseness, Taruma linelikeness, Taruma regularity, and Taruma roughness.

This study aims to develop robust regression models to predict sunspot numbers by incorporating independent features such as proton temperature, wind speed flow, proton density, and the Interplanetary Magnetic Field (IMF), along with exogenous topological variables derived from spectral solar images. These features, extracted from the SOHO dataset, provide crucial information obtained through advanced image processing techniques. Such techniques, relevant in computer vision and image processing, provide quantitative information about the structure, content, and visual characteristics of solar images, thus facilitating tasks like classification, detection, and segmentation (Tamura et al., 1978; Amadasun and King, 1989; Wu and Chen, 1992). We aim to determine the best approach for predicting sunspot numbers by comparing the performance of different machine learning models using these diverse features obtained from various SOHO spectral filters.

Our dataset supports the training of five machine learning regression models: Random Forest Regressor, Gradient Boosting Regressor, Extra Trees Regressor, Ada Boost Regressor, and Hist Gradient Boosting Regressor. By leveraging this dataset, we compare the effectiveness of these models in accurately predicting sunspot numbers. The objective is to evaluate how the inclusion of topological variables from SOHO images enhances the predictive accuracy of these machine learning models.

The primary objectives of this research are: (1) to assess the performance of various machine learning regression models in predicting sunspot numbers, considering both traditional solar parameters and novel topological attributes; and (2) to investigate the added value of incorporating these new variables derived from SOHO images in improving prediction accuracy. This study highlights the potential of advanced machine learning approaches to provide high-precision predictions of sunspot numbers, thereby offering deeper insights into the dynamics of solar activity.

While Mahdi et al. (2019) focused on predicting CME initiation using traditional classification algorithms based on sunspot number, we propose a research takes a different approach by incorporating unique topological features derived from solar images. These topological features, not considered in previous studies, capture intricate patterns of solar activity, providing additional context and potentially enhancing the predictive accuracy of our regression models for sunspot numbers.

Additionally, while XGBoost-DL (Dang et al., 2022) focuses on the power of deep learning and ensemble techniques, our study takes a different approach by incorporating these unique topological variables derived from solar images.

2. Data, methods and techniques

2.1. Data acquisition, preparation and description

The data used in this study come from three primary sources: OMNIWEB (https://omniweb.gsfc.nasa.gov/ow.html), the Royal Observatory of Belgium's Solar Influences Data Analysis Center (SILSO) (https://www.sidc.be/SILSO/datafiles), and a dataset from Mendeley Data and Data in Brief (https://data.mendeley.com/datasets/5gh3xbvc92/1).

The dataset from OMNIWEB provides four main variables, including wind flow speed (km/sec), proton density (n/cc), proton temperature (Kelvin), and Interplanetary Magnetic Field (IMF) (nT). The Brussels Observatory dataset offers corresponding sunspot values with daily resolution. Meanwhile, the Mendeley Data and Data in Brief dataset provides 14 topological and spectral variables. For a detailed description of how the data are constructed, refer to Sierra Porta and Tarazona-Alvarado (2023), Tarazona-Alvarado and Sierra-Porta (2023). These variables are generated from SOHO images taken with different solar filters, such as HMIIGR, HMIMAG (Helioseismic and Magnetic Imager Intensitygram), and EIT171, EIT195, EIT284, and EIT304 (Extreme Ultraviolet Imaging Telescope).

For the new variables and features, we curated a diverse and comprehensive dataset derived from time series, encompassing data from six distinct filters of the SOHO cameras. These datasets cover various resolutions and frequencies, providing a detailed and temporally rich view of solar activity. Noteworthy are the HMIIGR and HMIMAG filters, which operate in the visible and near-infrared spectrum, respectively, capturing images every hour and a half (Schou et al., 2012; Scherrer et al., 2012). Additionally, the EIT filters (EIT171, EIT195, EIT284, and

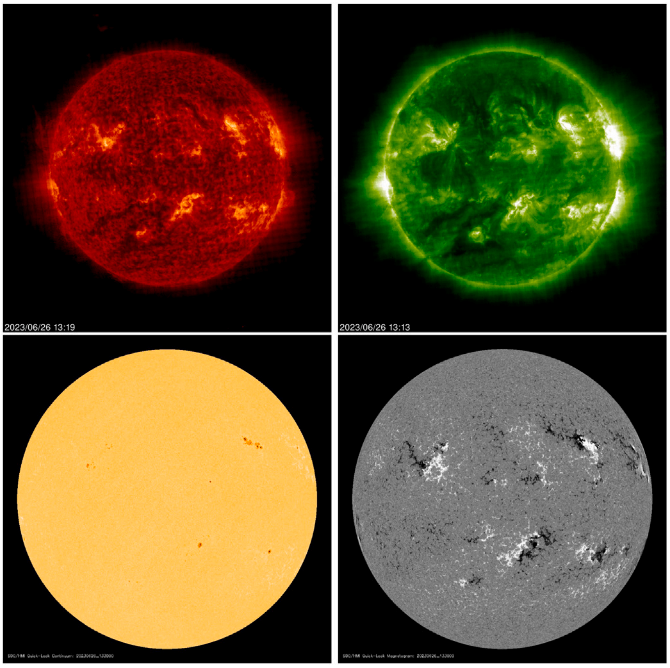

2023/06/26 13:19
2023/06/26 13:13
SDO/HMI Quick-Look Continuum: 20230626_133000
SDO/HMI Quick-Look Magnetogram: 20230626_133000

Fig. 1. Examples of solar images analyzed in our study as of June 26, 2023. The top image shows the Sun captured by the SOHO EIT304 filter (https://soho.nascom.nasa.gov/data/REPROCESSING/Completed/2023/eit304/20230626/20230626_1319_eit304_1024.jpg) and the SOHO EIT195 filter (https://soho.nascom.nasa.gov/data/REPROCESSING/Completed/2023/eit195/20230626/20230626_1313_eit195_1024.jpg), which highlight the extreme ultraviolet emission at wavelengths of 304 Å and 195 Å, respectively, revealing the structure of the solar chromosphere, transition region, hot corona, and solar flares. Additionally, bottom images include samples from the HMIIGR (https://soho.nascom.nasa.gov/data/REPROCESSING/Completed/2023/hmiigr/20230626/20230626_1330_hmiigr_1024.jpg) and HMIMAG (https://soho.nascom.nasa.gov/data/REPROCESSING/Completed/2023/hmimag/20230626/20230626_1330_hmimag_1024.jpg) filters, which capture intensitygrams and magnetograms of the solar surface, providing essential data for topological feature extraction. These images are crucial for deriving the topological features used in our machine learning models to predict sunspot numbers.

EIT304) in the extreme ultraviolet region capture two images per day at different wavelengths (Delaboudiniere et al., 1995; Kohl et al., 1995) (see Fig. 1).

This wide range of capture frequencies facilitates an in-depth exploration of solar activity patterns across various temporal scales. Furthermore, the computation of 14 key parameters in the images — including entropy, mean intensity, standard deviation, skewness, kurtosis, relative smoothness, uniformity, fractal dimension, and various Taruma metrics — adds significant depth to our dataset (Tamura et al., 1978; Amadasun and King, 1989; Wu and Chen, 1992). These parameters offer a comprehensive characterization of solar image properties, encompassing texture, complexity, and regularity.

Entropy represents the randomness and diversity of the image, calculated using Shannon's entropy formula. Mean intensity is the average of all pixel intensities in the image, providing an overall measure of brightness. Standard deviation indicates the dispersion of pixel intensities with respect to the mean intensity. Skewness measures the symmetry of the intensity distribution in the image. Kurtosis quantifies the shape of the intensity distribution, indicating the peakedness. Uniformity indicates how evenly the pixel intensities are distributed in the image. Relative smoothness provides a measure of the smoothness or roughness of the image. Taruma Contrast evaluates the contrast characteristics of the image based on standard deviation and kurtosis. Taruma Directionality quantifies the predominant direction of features within the image. Taruma Coarseness measures the texture or granularity of the image based on the coefficients of the wavelet transform. Taruma Linelikeness measures the presence and prevalence of linear patterns in the image. Taruma Regularity quantifies the uniformity and repeatability of patterns present in the image. Taruma Roughness evaluates the degree of irregularity or rough texture present in the image.

The data mining process (see Fig. 2) involves organizing, cleaning, and arranging the three different datasets. This process includes eliminating missing data and merging the datasets based on the temporal alignment of events. The sunspot resolution is daily, so daily averages of the heliospheric dynamics data are used, and the topological characteristics data are also resampled using Exponential Weighted Moving functions (EWM) method, which are useful for smoothing data and emphasizing more on recent observations. From the above, six datasets (one per filter in SOHO images) are produced with the same characteristics and variables for each type of filter wavelength used to obtain the SOHO images, covering the period from the beginning of 2011 to mid-2023.

To ensure the robustness of the predictive models, a thorough collinearity analysis was conducted. This analysis identified significant correlations between variables, which were then used to guide feature selection and engineering processes. The details and results of the correlation analysis are presented in the Results section.

For the moment, Fig. 3 illustrates the temporal behavior of the sunspot number (upper-left panel) along with various variables related to heliosphere dynamics and topological features. Specifically, it shows the time series for sunspot number, entropy, fractal dimension, Taruma's coarseness, Taruma's uniformity, bulk flow speed, proton density, proton temperature, and field magnetic average (FMA). These variables provide insights into the intricate dynamics of solar activity over the observed period.

After this process we obtain 6 datasets (one for each type of image analyzed using a different filter according to the SOHO, that is: HMIIGR, HMIMAG, EIT171, EIT195, EIT284 and EIT304), each of these datasets contains 6 heliospheric variables, 14 topological variables and the sunspot number, which will be our predicted variable.

2.2. Methods: Machine learning regressors

The main objective of machine learning (ML) techniques is to develop models capable of performing tasks such as prediction or estimation. In regression, the goal is to predict continuous values rather than classify data into predefined classes. When developing regression models using ML techniques, both training errors (errors on the training data) and generalization errors (expected errors on the testing data) can occur. A good regression model should fit the training set well and accurately predict new, unseen data. Overfitting, where test error rates increase while training error rates decrease, is related to model complexity and should be minimized to achieve the lowest generalization error. The bias–variance decomposition method formally analyzes the expected generalization error by measuring the bias (error rate) and variance (sensitivity to training set fluctuations) components, with the overall expected error being the sum of both.

The primary objective of our study is to enhance the predictive accuracy of sunspot numbers by incorporating both heliospheric and topological features derived from solar images. Given the complex and dynamic nature of solar activity, it is essential to employ robust machine learning methodologies that can effectively capture and model these intricacies. Traditional approaches relying solely on heliospheric variables may fall short in capturing the full spectrum of influences on sunspot numbers. Therefore, our approach leverages advanced machine learning algorithms and a comprehensive dataset that includes additional topological features from SOHO images. By comparing models that use only heliospheric variables with those that also incorporate topological features, we aim to demonstrate the added predictive power of these features. This dual-scenario approach, applied across six different types of solar images, allows for a thorough

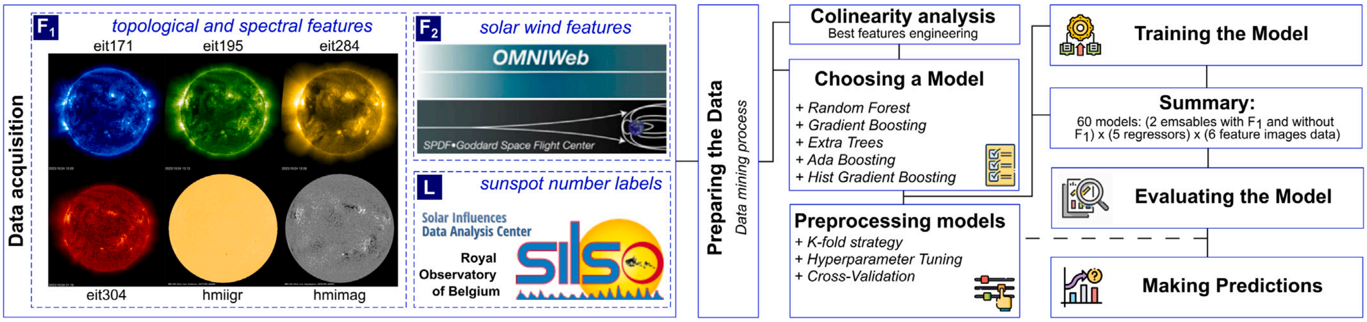


Fig. 2. Overview of the methodological framework for data acquisition, preparation, and model training. The process begins with data acquisition, involving the collection of topological and spectral features from solar images ($\mathbf{F_1}$) and solar wind features from the OMNIWeb database ($\mathbf{F_2}$). Sunspot number labels are sourced from the SILSO database (**L**). The data preparation phase includes preprocessing steps such as collinearity analysis, feature engineering, and the selection of appropriate machine learning models.

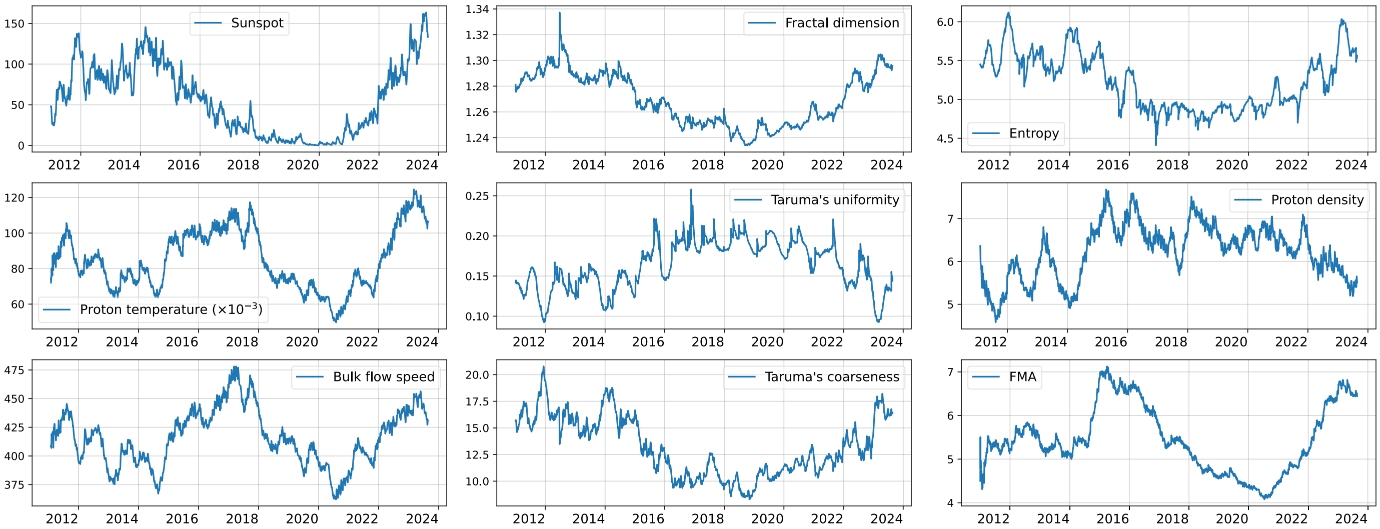


Fig. 3. Temporal behavior of the sunspot number (upper-left panel), as well as various variables related to heliosphere dynamics and topological features calculated from the SOHO EIT171 filter images.

evaluation of the effectiveness of topological information in improving sunspot prediction models.

To develop regression models, we created a total of 60 models. These consist of 5 regressors × 6 image types × 2 scenarios. The two scenarios include: one using only heliospheric variables, and another using both heliospheric and topological features derived from SOHO images. This approach allows us to compare the predictive power of models and demonstrate how topological features can enhance traditional models.

The data was first split into training and testing sets using the k-fold cross-validation method (Schaffer, 1993; Tougui et al., 2021; Ramezan et al., 2019) and hyperparameter tuning (Yang and Shami, 2020; Weerts et al., 2020). The training set was used to train the models, while the testing set was used to evaluate their performance.

The machine learning models used in this work are the following. The Random Forest Regressor (RF) (Breiman, 2001) is an ensemble method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees, thereby reducing variance and overfitting. This robustness makes RF suitable for handling the noisy, high-dimensional nature and nonlinear behavior of solar activity data. The 11-year solar cycle leads to periodic variations in sunspot numbers, which often exhibit near-normal distributions but with noticeable skewness and asymmetry. The Gradient Boosting Regressor (GB) (Friedman, 2001) builds an ensemble of weak prediction models sequentially, with each new tree correcting the errors of the previous ones. This approach is particularly effective for capturing the complex, nonlinear relationships in solar activity data. Similarly, the Extra Trees Regressor (ET) (Geurts et al., 2006) introduces additional randomness during tree construction, improving model diversity and performance in datasets with small or noisy characteristics, such as those found in sunspot observations.

The AdaBoost Regressor (AB) (Freund and Schapire, 1997) enhances prediction accuracy by focusing on reducing variance and assigning more weight to incorrectly predicted instances in each iteration. This makes AB resistant to overfitting and suitable for the complex distributions observed in sunspot data. Lastly, the Extreme Gradient Boosting Regressor (XGB) (Chen and Guestrin, 2016) sequentially trains weak models to correct previous errors, optimizing a specific loss function to create robust predictions. XGB's flexibility and ability to handle high-dimensional data with numerous features make it particularly effective for capturing the intricacies of sunspot number variations over the 11-year solar cycle, accommodating the inherent skewness and asymmetry.

For the RF, GB, ET, AB, and XGB models, different values for the number of trees (N_ESTIMATORS), learning rate (LEARNING_RATE), and maximum tree depth (MAX_DEPTH) were explored using GridSearchCV and hyperparameter tuning with a cross-validation strategy (Schaffer, 1993). The parameters evaluated included MAX_DEPTH: [5, 10, 20, None], MAX_FEATURES: [5, 10, 20, 'auto', 'sqrt', 'log2', None], N_ESTIMATORS: [30, 50, 100, 200, 300], LEARNING_RATE: [0.01, 0.05, 0.1, 0.2, 0.5], MAX_ITER: [50, 100, 150].

The cross-validation strategy was used to select the best models in each case, allowing evaluation of each model's performance on an independent validation set and optimization of its parameters through grid search. Once the best estimators and parameters were determined for each model, their performance was evaluated on the independent test set.

For the design, implementation, and application of our regression models, each dataset was split into three distinct subsets: a training set, a testing set, and a validation set. The training set comprises 70% of the complete data, selected entirely at random. The testing set consists of the remaining 30% of the data. Additionally, a third validation set was created by randomly selecting 70% of the original data, using a different seed to ensure that this set includes elements from both the training and testing sets. This approach allows for robust model evaluation and ensures that the models are not overfitted to a particular subset of data, providing a comprehensive assessment of their predictive performance.

2.3. Performance metrics

Once a regression model is obtained using one or more ML techniques, it is important to estimate the model's performance. The performance analysis of each proposed model is measured in terms of r2-score (r2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Explained Variance Score (EVS), and Mean Tweedie Deviance (MTD).

The r2-score measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It ranges from 0 to 1, with 1 indicating a perfect fit between the predicted and actual values. A value of 0 indicates that the model does not explain any of the variability in the data. RMSE is a metric that measures the average distance between the predicted and actual values of the dependent variable. It is calculated by taking the square root of the mean of the squared differences between the predicted and actual values.

The EVS measures the proportion of the variance in the dependent variable explained by the independent variable(s). It ranges from 0 to 1, with 1 indicating a perfect fit between the predicted and actual values. A value of 0 indicates that the model does not explain any of the variability in the data. The MTD metric measures the deviation of the predicted values from the true values of the dependent variable. It is calculated by taking the mean of the Tweedie deviances, which measure the difference between the predicted and actual values. In regression model evaluation, a lower MTD value indicates better model performance, as it signifies that the predictions are closer to the true values. However, the optimal MTD value may vary depending on the specific context and application.

The predictive accuracy of the model is computed from the testing set, which provides an estimation of the generalization errors. To obtain reliable results regarding the predictive performance of a regression model, it is crucial that training and testing samples are sufficiently large and independent, with known labels for the testing sets. Common methods for evaluating the performance of a regression model by splitting the initial labeled data into subsets include: (i) Holdout Method, (ii) Random Sampling, (iii) Cross-Validation, and (iv) Bootstrap.

In our study, we employed the random sampling method with a bootstrap option. This approach is similar to the Holdout method, where data samples are partitioned into two separate sets: the training set and the test set. However, random sampling involves repeating this partitioning process multiple times, with training and test instances selected randomly each time to better estimate accuracy. By using the bootstrap option, samples are selected with replacement, meaning that after being chosen for training, they are returned to the entire data set. This method allows for a comprehensive evaluation of the model's predictive performance by ensuring that the model is tested on a variety of data subsets.

Table 1
Summary of variables retained after application of the collinearity between variables removal algorithm. The N_T column refers to the number of topological variables retained from the initial 14 totals, while the $R\%$ column refers to the percentage reduction of variables with respect to the original (20).

Dataset	Keep variables	N_T	$R\%$
HMIIGR	FMA, Bulk Flow speed, Proton Density, Temperature, entropy, standard deviation, fractal dimension, Taruma regularity	4	60
HMIMAG	FMA, Bulk Flow speed, Proton Density, Temperature, entropy, standard deviation, fractal dimension, Taruma roughness	4	60
EIT171	FMA, Bulk Flow speed, Proton Density, Temperature, standard deviation, relative smoothness, fractal dimension, Taruma directionality, Taruma linelikeness, Taruma regularity, Taruma roughness	7	45
EIT195	FMA, Bulk Flow speed, Proton Density, Temperature, standard deviation, kurtosis, relative smoothness, fractal dimension, Taruma roughness, Taruma coarseness, Taruma directionality	7	45
EIT284	FMA, Bulk Flow speed, Proton Density, Temperature, standard deviation, skewness, fractal dimension, Taruma directionality, Taruma linelikeness, Taruma regularity, Taruma roughness	7	45
EIT304	FMA, Bulk Flow speed, Proton Density, Temperature, relative smoothness, fractal dimension, Taruma directionality, Taruma regularity, Taruma linelikeness, Taruma roughness	6	50

3. Results of ML applications in sunspot

In regression analysis, multicollinearity among independent variables can lead to unstable estimates and reduce the interpretability of the model. To address this issue, we implemented a strategy to identify and remove highly collinear variables based on a specified correlation threshold. This process involves calculating the correlation matrix for all independent variables and excluding those with correlation coefficients exceeding the threshold of 0.7. By doing so, we aim to ensure that the remaining variables provide unique and non-redundant information, improving the model's stability and predictive power. In this sense, some Python libraries (https://pypi.org/project/collinearity/) were used for this task.

For instance, Fig. 4 presents a heatmap of the correlation matrix for the EIT171 dataset, illustrating the relationships between all variables before collinearity removal.

The initial dataset comprised 20 independent variables: 4 heliospheric variables (FMA, Bulk Flow speed, Proton Density, Temperature) and 14 topological variables (entropy, mean intensity, standard deviation, skewness, kurtosis, relative smoothness, uniformity, fractal dimension, Taruma contrast, Taruma directionality, Taruma coarseness, Taruma linelikeness, Taruma regularity, Taruma roughness). The target variable to be predicted was the sunspot number (SSN).

After applying the collinearity removal process with a threshold of 0.7, the variables retained for each dataset were significantly reduced. The Table 1 shows the variables that have been retained after elimination due to collinearity between variables.

Table 2 presents the results of the model evaluation metrics for the various datasets. The numbers inside parentheses in each cell represent the metrics for the sunspot regression using flow speed, temperature, density, and IMF as predictor variables, along with the 14 topological

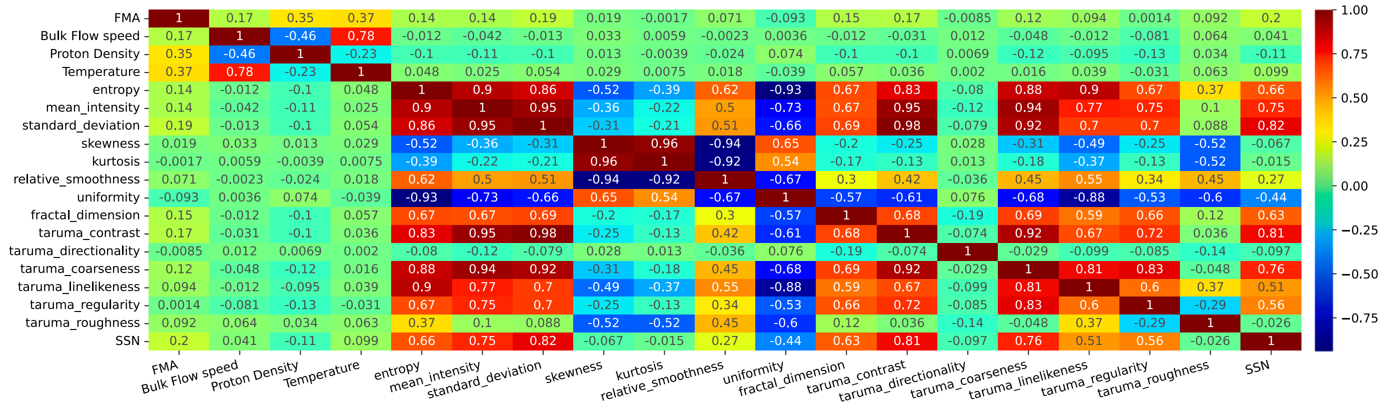


Fig. 4. Heatmap of the correlation matrix for the EIT171 dataset. Highly correlated variables (correlation coefficient greater than 0.7) were removed to mitigate multicollinearity.

Table 2
Performance metrics of the five machine learning algorithms for predicting sunspot number based on the six datasets with the same characteristics and variables for each of the types of filter wavelengths to obtain the SOHO images. The metrics include r2-score, MAE, RMSE and MTD, and are shown for each of the six datasets: EIT171, EIT195, EIT284, EIT304, HMIIGR, and HMIMAG. The values in parentheses represent the correspondent metric for scenario 2, that is, including topological variables in comparison with scenario 1 (only heliospheric variables, without parentheses). The best performance for each metric is highlighted in bold.

Metric	Regr.	EIT171	EIT195	EIT284	EIT304	HMIIGR	HMIMAG
r2-score	AB	0.11(0.74)	0.12(0.74)	0.11(0.78)	0.12(0.68)	0.12(0.42)	0.12(0.82)
	ET	**0.31(0.95)**	**0.31(0.95)**	**0.31(0.96)**	**0.31(0.96)**	**0.32(0.94)**	**0.32(0.97)**
	GB	0.25(0.85)	0.25(0.86)	0.25(0.91)	0.25(0.93)	0.25(0.93)	0.25(0.93)
	RF	0.19(0.93)	0.19(0.93)	0.19(0.94)	0.19(0.94)	0.19(0.89)	0.19(0.95)
	XGB	0.24(0.87)	0.24(0.91)	0.24(0.93)	0.24(0.94)	0.24(0.93)	0.24(0.92)
MAE	AB	38.83(19.84)	38.99(19.44)	39.02(18.01)	38.76(21.79)	39.2(31.76)	39.2(17.62)
	ET	**33.53(3.97)**	**33.58(3.79)**	**33.67(3.28)**	**33.55(3.32)**	**33.94(5.12)**	**33.94(3.38)**
	GB	35.26(14.26)	35.26(13.64)	35.04(10.99)	35.18(8.11)	35.28(5.95)	35.28(11.48)
	RF	36.77(8.59)	36.74(8.23)	36.68(7.3)	36.82(7.47)	36.89(10.1)	36.89(7.41)
	XGB	35.4(12.81)	35.31(10.73)	35.18(8.42)	35.37(7.48)	35.43(5.86)	35.43(9.85)
RMSE	AB	46.92(25.45)	46.8(25.38)	46.89(23.41)	46.87(28.41)	46.81(38.0)	46.81(22.08)
	ET	**41.36(11.04)**	**41.42(10.75)**	**41.52(9.5)**	**41.31(9.7)**	**41.69(12.52)**	**41.69(9.14)**
	GB	43.24(19.36)	43.3(18.72)	43.07(15.51)	43.13(12.79)	43.12(13.51)	43.12(15.5)
	RF	44.83(13.36)	44.89(13.03)	44.78(11.94)	44.94(12.22)	44.89(15.49)	44.89(11.25)
	XGB	43.44(17.65)	43.43(15.5)	43.32(13.02)	43.4(12.49)	43.36(13.37)	43.36(13.71)
MTD	AB	2201.73(647.88)	2190.15(644.32)	2199.09(548.1)	2196.37(806.89)	2191.3(1443.75)	2191.3(487.56)
	ET	**1710.36(121.85)**	**1715.53(115.57)**	**1723.62(90.2)**	**1706.13(94.06)**	**1738.46(156.87)**	**1738.46(83.6)**
	GB	1870.07(374.91)	1874.81(350.37)	1854.97(240.69)	1860.33(163.51)	1859.64(182.47)	1859.64(240.18)
	RF	2010.02(178.6)	2015.09(169.74)	2005.46(142.64)	2019.29(149.31)	2015.42(239.8)	2015.42(126.67)
	XGB	1887.13(311.52)	1885.74(240.37)	1877.06(169.51)	1883.5(155.89)	1880.17(178.79)	1880.17(187.99)

variables. In contrast, the numbers without parentheses reflect the metrics when using only flow speed, temperature, density, and IMF as predictor variables.

The results in Table 2 clearly indicate that the models incorporating the remaining topological variables after collinearity analysis, perform better than those using only the heliospheric variables. Additionally, the Extra Trees Regressor and Random Forest Regressor models outperform the other models evaluated for all datasets considered.

After a thorough analysis and taking into account the results summarized in Table 2, it is found that ETR is the best regression model that best predicts sunspots for all considered dataset. For the case of the HMIIGR image dataset, it was determined that the most important variables contributing to the best result are: entropy (27.04%), standard deviation (24.14%), Taruma regularity (23.03%), fractal dimension (8.59%), and FMA (5.56%), covering up to 88% of the total importance of the model. In this case, we see that only one of the heliospheric variables contributes minimally. However, the power of the new model lies in the inclusion of new topological variables.

Similarly, the best regression model using HMIMAG image features dataset determined that the most important variables are: fractal dimension (40.98%), standard deviation (23.99%), entropy (19.41%), and Taruma roughness (6.41%). These variables cover up to 91% of the total importance of the model. This result suggests that topological features extracted from HMIMAG images significantly impact model performance, providing new insights and potential for improved prediction accuracy compared to previous models using only heliospheric variables.

A similar conclusion is reached in the analyses performed on other types of images, such as EIT171, 195, 284, and 304. It is observed that heliospheric variables, obtained from the original OMNIWEB data, account for a minimal proportion of the model's significance, generally less than 10%. This finding reinforces the idea that features extracted from image topology are critical for improving predictive models. The consistency in these results suggests that the inclusion of new topological variables brings significant improvement in the predictive ability of the models, regardless of the type of image analyzed. A complete importance variable for all datasets are shown in Table 3.

In our study, we found that the use of topological variables significantly improved the performance of the machine learning models. Specifically, we observed that the average r2-score for all models generated without the use of topological variables was approximately 0.31, while the inclusion of topological variables increased this r2-score to over 0.94. This indicates a substantial improvement in the model performance when topological variables are included. For clarity, the reported averages are calculated across all models and datasets. Additionally, the RMSE of the model decreased from 41.5 average to approximately 10.4 average when topological variables were included. Finally, we found that the Extra Trees Regressor (ET) was the best

Table 3
Percentage of importance of the variables that contribute most to the ETR model for all the datasets considered in this study. The first 8 most important variables are shown in order of highest contribution.

HMIIGR	Importance (%)	HMIMAG	Importance (%)	EIT171	Importance (%)
entropy	27.04	fractal_dimension	40.98	standard_deviation	34.73
standard_deviation	24.14	standard_deviation	23.99	relative_smoothness	15.27
taruma_regularity	23.03	entropy	19.41	taruma_directionality	13.26
fractal_dimension	8.59	taruma_roughness	6.41	taruma_regularity	10.04
FMA	5.56	FMA	2.50	fractal_dimension	8.09
Proton Density	4.21	Proton Density	2.40	taruma_linelikeness	6.61
Bulk Flow speed	4.09	Bulk Flow speed	2.22	taruma_roughness	3.65
Temperature	3.31	Temperature	2.04	FMA	2.35
EIT195	**Importance (%)**	**EIT284**	**Importance (%)**	**Variable**	**Importance (%)**
standard_deviation	34.47	standard_deviation	54.10	relative_smoothness	49.26
relative_smoothness	21.74	taruma_linelikeness	12.44	taruma_regularity	11.74
taruma_coarseness	19.27	skewness	7.57	fractal_dimension	9.98
taruma_roughness	4.70	fractal_dimension	6.96	taruma_roughness	8.23
kurtosis	4.34	taruma_regularity	5.44	taruma_linelikeness	6.81
taruma_directionality	3.88	taruma_roughness	4.40	taruma_directionality	5.22
fractal_dimension	3.06	taruma_directionality	2.79	Bulk Flow speed	2.33
FMA	2.56	Bulk Flow speed	1.63	FMA	2.32

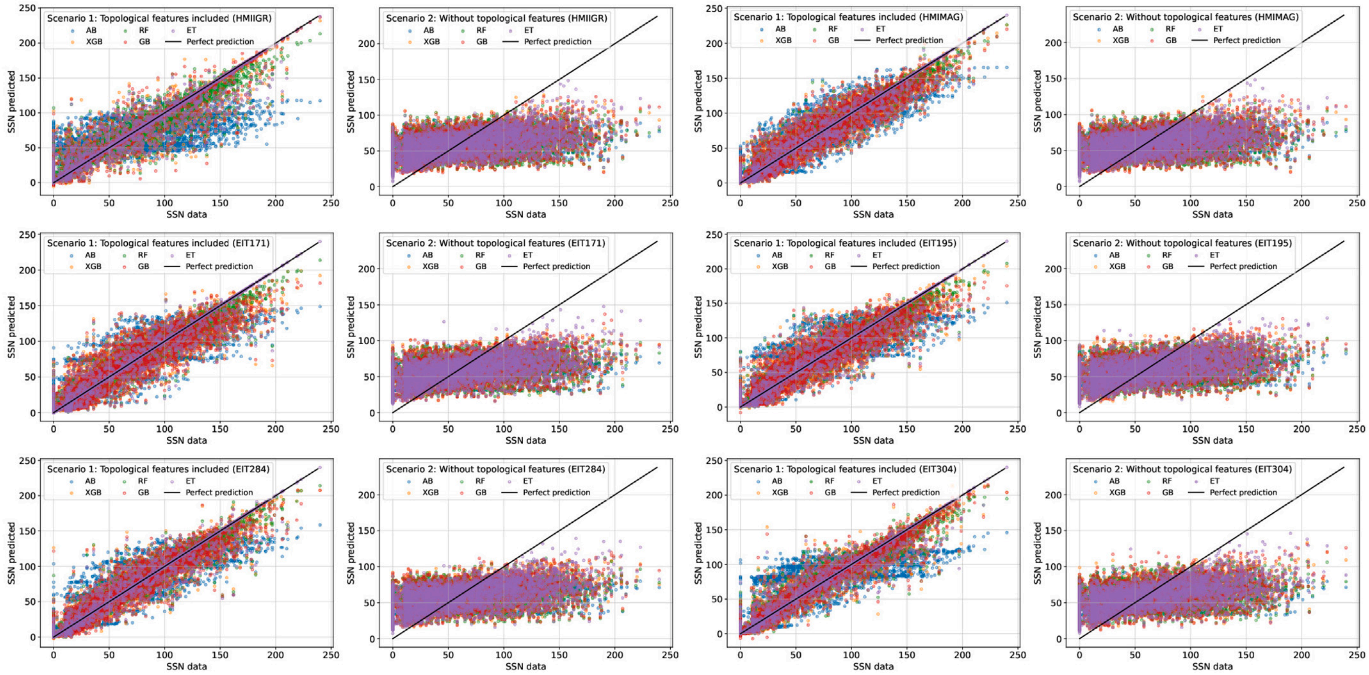


Fig. 5. Visual inspection of predicted SSN versus SSN with original measured data. Each panel shows the comparison of data and prediction using only OMNI variables (right) and also using topological variables (left) for each type of imagery used.

model among all the models considered in this study, based on all the control metrics used.

Fig. 5 visually presents the SSN prediction performance of all the models and the data obtained from each type of filter used on the Sun in the SOHO images. The ET model, as shown in Table 2, is confirmed to be the best regression model, demonstrating superior performance and results. It is evident that when topological variables are not used, the models fail to adequately predict the sunspot number. However, with the inclusion of topological variables, the predictions are well-adjusted and accurate.

Additionally, we can observe that the ET model outperforms all other models for both high and low sunspot numbers. However, it is also noted that for all models, there is higher dispersion at lower SSN values and very high prediction accuracy for higher SSN values.

In other word, ETR model consistently achieves superior metrics compared to other regressors, particularly when topological variables are included. This conclusion is further reinforced by examining the disaggregated root mean squared error metrics for the training, test, total, and validation sets (See Table 4).

The inclusion of topological variables not only improves prediction accuracy but also demonstrates no evident overfitting. For example, in the HMIIGR dataset, the training set and testing set have approximately the same RMSE, both in two scenarios with and without topological variables. This indicates that the model generalizes well to unseen data. The validation set RMSE of approximately value further supports the robustness of the model. Similar trends are observed in other datasets, where the validation set MSE remains low and comparable to the training and test set MSE values.

This study demonstrates that the inclusion of topological variables from solar images significantly enhances the performance of sunspot prediction models. While the XGBoost-DL model proposed by other researchers (Dang et al., 2022) uses a two-level nonlinear ensemble method to combine deep learning models, achieving an RMSE of 25.70 and an MAE of 19.82, our approach with the Extra Trees Regressor (ET) and other models incorporates topological features. These features

Table 4
Disaggregated root mean squared error metric for the ET model across different datasets and evaluation sets.

Scenario	Dataset	HMIIGR	HMIMAG	EIT171	EIT195	EIT284	EIT304
Only Heliospheric variables	Training	41.39	41.39	40.87	41.15	40.84	41.25
	Testing	42.61	42.61	43.03	41.98	42.67	42.29
	All	41.69	41.69	41.42	41.36	41.31	41.52
	Validation	41.95	41.95	41.21	41.26	41.11	41.41
Heliospheric + Topological variables	Training	8.93	12.5	10.48	10.85	9.17	9.52
	Testing	9.76	12.59	11.52	11.59	10.42	10.21
	All	9.14	12.52	10.75	11.04	9.5	9.7
	Validation	9.29	12.77	10.66	10.99	9.38	9.58

capture detailed solar activity patterns, which heliospheric variables alone might miss. The key difference in our study lies in the use of these unique topological variables, which provide additional context and improve model accuracy by accounting for the complex structures observed in solar images.

Incorporating topological variables such as entropy into regression prediction models can significantly enhance their predictive capabilities by capturing the underlying complexity and randomness within the data. Entropy measures the degree of disorder or uncertainty in a dataset, making it a valuable predictor for models dealing with dynamic systems like solar activity. For example, Shannon entropy has been effectively used in molecular property predictions (Guha and Velegol, 2023), demonstrating that entropy-based descriptors can reduce prediction errors and improve model accuracy by capturing intricate data patterns. This suggests that entropy can similarly enhance solar activity models by accounting for the chaotic and unpredictable nature of solar phenomena.

Fractal dimension is another topological feature that quantifies the complexity of structures within an image, offering a deeper understanding of spatial patterns. Fractal analysis has been widely used in various scientific fields to identify self-similar patterns that traditional metrics might miss. For instance, fractal dimensions have been used to improve the prediction accuracy of molecular properties in cheminformatics, highlighting their potential to capture complex relationships within data (Kozak and Juszczuk, 2023). By integrating fractal dimension into solar activity models, we can better represent the intricate structures of sunspots and other solar features, leading to more precise predictions.

Taruma features, including contrast, directionality, coarseness, linelikeness, regularity, and roughness, provide a detailed characterization of texture within images, capturing fine details and variations essential for understanding solar dynamics. These features have been proven effective in various machine learning applications, such as texture classification and image analysis (Aggarwal and Kumar, 2021; Salem and Abdelkrim, 2020). In the context of solar image analysis, Taruma features can highlight critical patterns and structures that correlate with solar activity, improving the model's ability to predict sunspot numbers accurately.

Overall, the integration of these topological variables enhances the robustness and accuracy of regression prediction models by providing a richer, multi-dimensional view of the data. Studies have shown that combining multiple types of descriptors, such as Shannon entropy and fractal dimensions, can significantly improve the performance of machine learning models in different applications (Kozak and Juszczuk, 2023; Saroughi et al., 2024; Keller, 2019; Guha and Velegol, 2023). By incorporating these advanced features, our models achieve higher predictive accuracy, as evidenced by the substantial improvement in r2-scores and reduction in mean squared errors. This approach not only advances our understanding of solar dynamics but also sets a new standard for predictive modeling in space weather forecasting, highlighting the potential for further improvements through the use of sophisticated data descriptors. This indicates that these topological features capture essential non-linear patterns and relationships in the solar images, which are not adequately represented by heliospheric variables alone.

Future work should explore the application of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to further enhance the prediction capabilities. These models are well-suited to capture complex spatial and temporal dependencies in data, respectively. For instance, CNNs can be employed to automatically extract high-level features from solar images, while RNNs or Long Short-Term Memory (LSTM) networks can model temporal sequences of sunspot activity, potentially leading to even more accurate and robust forecasts (Chollet, 2021). Advanced Feature Engineering

Another promising direction is to refine and expand the set of topological features. Higher-order statistics, advanced texture descriptors like Local Binary Patterns (LBP), and wavelet transform coefficients could provide additional valuable insights. These features can capture more intricate details of the solar images' texture and structure, contributing to a more comprehensive understanding of solar dynamics. Research in image processing and computer vision suggests that these advanced features often enhance the performance of machine learning models in various applications (Chollet, 2021; He et al., 2016; Guo et al., 2022).

Developing hybrid models that combine different machine learning techniques can also be beneficial. For instance, ensemble methods that integrate predictions from traditional machine learning models and deep learning models could leverage the strengths of both approaches. Techniques like stacking or blending, where multiple models are trained and their outputs combined, could provide a more robust and accurate prediction system. This approach has been shown to improve model performance in various predictive tasks (De Alwis and Samadi, 2024). Temporal Dynamics and Sequence Modeling

Incorporating models designed to handle temporal dynamics, such as Temporal Convolutional Networks (TCNs) or Transformer models, could further improve the predictive accuracy of sunspot numbers. These models are particularly adept at capturing long-term dependencies and trends, which are critical in forecasting sunspot activity. The Transformer model, with its attention mechanisms, has been especially successful in various sequence modeling tasks and could be adapted for sunspot prediction.

Moreover, improving the interpretability of the models is crucial for gaining insights into the underlying physical processes of solar activity. Techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can help elucidate the contributions of individual features to the model's predictions. This understanding can guide further feature engineering efforts and provide valuable information to solar physicists (Wang et al., 2021).

Lastly, integrating topological variables with other types of solar data, such as solar flare occurrences and coronal mass ejections, could provide a more holistic view of solar activity. This multimodal approach can enhance the models' ability to predict not only sunspot numbers but also other related phenomena, thereby contributing to a comprehensive space weather forecasting system (Benz, 2017).

4. Conclusions and future directions

Based on the findings of this study, the following conclusions can be drawn:

The study demonstrates that the inclusion of topological variables, in addition to the four primary heliospheric variables (wind flow speed, proton density, proton temperature, and IMF), significantly enhances the performance of regression models. In general, the r2-score improves from an average of 0.40 to over 0.90, and the variance explained by the models increases from approximately 0.3 to 0.93. The most important variables for the regression models include IMF, flow speed, proton density, proton temperature, entropy, standard deviation, fractal dimension, Taruma roughness, and Taruma regularity for HMIIGR and HMIMAG images. For EIT171, EIT195, EIT284, and EIT304 images, the significant variables also include relative smoothness, Taruma directionality, Taruma linelikeness, kurtosis, and Taruma coarseness.

The Extra Trees Regressor (ET) emerges as the best model among all those evaluated in this study, achieving an average r2-score of over 0.97 for nearly all models and variables across all types of supporting images.

The evaluation of model performance was based on four metrics: r2-score, root mean squared error (RMSE), explained variance score (EVS), and mean Tweedie deviance (MTD). Cross-validation and grid search were employed for hyperparameter optimization. The results indicate that model performance varies depending on the type of data used, with certain models performing better than others for specific datasets.

In conclusion, this study provides valuable insights into the use of machine learning algorithms for predicting sunspot numbers based on heliospheric dynamics data and topological and spectral variables. The results confirm that the inclusion of topological variables significantly improves model performance, with the Extra Trees Regressor being the most effective model among those considered. The study also emphasizes the importance of using appropriate evaluation metrics and hyperparameter optimization techniques to develop accurate and reliable models.

Future research directions include exploring the use of Long Short-Term Memory (LSTM) networks for sunspot forecasting and prediction. The authors plan to use sunspot time series data to train and test LSTM network models. The objective is to achieve a deeper understanding of sunspot forecasting and to evaluate the effectiveness of LSTM networks in this context. The process will involve the general steps of time series analysis, data preparation, and model training using appropriate LSTM network configurations.

CRediT authorship contribution statement

D. Sierra-Porta: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **M. Tarazona-Alvarado:** Writing – original draft, Software, Formal analysis, Data curation. **D.D. Herrera Acevedo:** Software, Methodology, Formal analysis, Data curation.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: David Sierra Porta reports equipment, drugs, or supplies was provided by Universidad Tecnológica de Bolívar. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is already in web and we share the links for dataset used in this paper.

Acknowledgments

DSP is grateful to the Universidad Tecnológica de Bolívar for the support received during the research process. While the research did not receive any financial contribution from internal or external funding sources, the Research Direction provided logistical support and computational equipment for the development of this study.

References

Aggarwal, A., Kumar, M., 2021. Image surface texture analysis and classification using deep learning. Multimedia Tools Appl. 80 (1), 1289–1309. doi:10.1007/s11042-020-09520-2.

Alexakis, P., Mavromichalaki, H., 2019. Statistical analysis of interplanetary coronal mass ejections and their geoeffectiveness during the solar cycles 23 and 24. Astrophys. Space Sci. 364 (11), 187. doi:10.1007/s10509-019-3677-y.

Amadasun, M., King, R., 1989. Textural features corresponding to textural properties. IEEE Trans. Syst. Man Cybern. 19 (5), 1264–1274. doi:10.1109/21.44046.

Barentine, J.C., 2022. Night sky brightness measurement, quality assessment and monitoring. Nat. Astron. 6 (10), 1120–1132. doi:10.1038/s41550-022-01756-2.

Benz, A.O., 2017. Flare observations. Living Rev. Solar Phys. 14, 1–59. doi:10.1007/s41116-016-0004-3.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324.

Burt, J., Smith, B., 2012. Deep space climate observatory: The DSCOVR mission. In: 2012 Ieee Aerospace Conference. IEEE, pp. 1–13. doi:10.1109/AERO.2012.6187025.

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. pp. 785–794. doi:10.1145/2939672.2939785.

Chollet, F., 2021. Deep learning with Python. Simon and Schuster, ISBN: 9781617294433.

Cliver, E.W., Pötzi, W., Veronig, A.M., 2022. Large sunspot groups and great magnetic storms: magnetic suppression of CMEs. Astrophys. J. 938 (2), 136. doi:10.3847/1538-4357/ac847d.

Dang, Y., Chen, Z., Li, H., Shu, H., 2022. A comparative study of non-deep learning, deep learning, and ensemble learning methods for sunspot number prediction. Appl. Artif. Intell. 36 (1), 2074129. doi:10.1080/08839514.2022.2074129.

Dani, T., Sulistiani, S., 2019. Prediction of maximum amplitude of solar cycle 25 using machine learning. In: J. Phys. Conf. Series. 1231, (1), IOP Publishing, 012022. doi:10.1088/1742-6596/1231/1/012022.

De Alwis, T.P., Samadi, S.Y., 2024. Stacking-based neural network for nonlinear time series analysis. Stat. Methods Appl. 1–24. doi:10.1007/s10260-024-00746-0.

Delaboudiniere, J.-P., Artzner, G., Brunaud, J., Gabriel, A.H., Hochedez, J.-F., Millier, F., Song, X., Au, B., Dere, K., Howard, R.A., et al., 1995. EIT: extreme-ultraviolet imaging telescope for the SOHO mission. SOHO Mission 291–312. doi:10.1007/978-94-009-0191-9_8.

Drews, A., Huo, W., Matthes, K., Kodera, K., Kruschke, T., 2022. The sun's role in decadal climate predictability in the north atlantic. Atmos. Chem. Phys. 22 (12), 7893–7904. doi:10.5194/acp-22-7893-2022.

Floyd, L., Tobiska, W.K., Cebula, R.P., 2002. Solar UV irradiance, its variation, and its relevance to the earth. Adv. Space Res. 29 (10), 1427–1440. doi:10.1016/S0273-1177(02)00202-8.

Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55 (1), 119–139. doi:10.1006/jcss.1997.1504.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 1189–1232, https://www.jstor.org/stable/2699986.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. doi:10.1007/s10994-006-6226-1.

Gour, P.S., Singh, N.P., Soni, S., Saini, S.M., 2021. Observation of coronal mass ejections in association with sun spot number and solar flares. In: IOP Conference Series: Materials Science and Engineering, Vol. 1120, No. 1. IOP Publishing, 012020. doi:10.1088/1757-899X/1120/1/012020.

Grauer, A.D., Grauer, P.A., 2021. Linking solar minimum, space weather, and night sky brightness. Sci. Rep. 11 (1), 23893. doi:10.1038/s41598-021-02365-1.

Guha, R., Velegol, D., 2023. Harnessing Shannon entropy-based descriptors in machine learning models to enhance the prediction accuracy of molecular properties. J. Cheminformat. 15 (1), 54. doi:10.1186/s13321-023-00712-0.

Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R.R., Cheng, M.-M., Hu, S.-M., 2022. Attention mechanisms in computer vision: A survey. Comput. Visual Media 8 (3), 331–368. doi:10.1007/s41095-022-0271-y.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. doi:10.1109/IEEESTD.1997.85951.

Hill, M., Allen, R., Kollmann, P., Brown, L., Decker, R., McNutt, R., Krimigis, S., Andrews, G., Bagenal, F., Clark, G., et al., 2020. Influence of solar disturbances on galactic cosmic rays in the solar wind, heliosheath, and local interstellar medium: Advanced composition explorer, new horizons, and voyager observations. Astrophys. J. 905 (1), 69. doi:10.3847/1538-4357/abb408.

Keller, K., 2019. Entropy measures for data analysis: Theory, algorithms and applications. Entropy 21 (10), 935. doi:10.3390/e23111496.

Khan, T., Arafat, F., Mojumdar, M.U., Rajbongshi, A., Siddiquee, S.M.T., Chakraborty, N.R., 2020. A machine learning approach for predicting the sunspot of solar cycle. In: 2020 11th International Conference on Computing, Communication and Networking Technologies. ICCCNT, IEEE, pp. 1–4. doi:10.1109/ICCCNT49239.2020.9225427.

Kohl, J.L., Esser, R., Gardner, L.D., Habbal, S., Daigneau, P.S., Dennis, E., Nystrom, G., Panasyuk, A., Raymond, J., Smith, P., et al., 1995. The ultraviolet coronagraph spectrometer for the solar and heliospheric observatory. SOHO Mission 313–356. doi:10.1007/978-94-009-0191-9_9.

Kozak, J., Juszczuk, P., 2023. Entropy in Real-World Datasets and Its Impact on Machine Learning. MDPI-Multidisciplinary Digital Publishing Institute, doi:10.3390/books978-3-0365-7849-1.

Le Mouël, J.-L., Lopes, F., Courtillot, V., 2019. A solar signature in many climate indices. J. Geophys. Res.: Atmos. 124 (5), 2600–2619. doi:10.1029/2018JD028939.

Li, Q., Wan, M., Zeng, S.-G., Zheng, S., Deng, L.-H., 2021. Predicting the 25th solar cycle using deep learning methods based on sunspot area data. Res. Astron. Astrophys. 21 (7), 184. doi:10.1088/1674-4527/21/7/184.

Mahdi, M.M., Tipu, M.A.N., Halder, C., Rahman, K.F., 2019. Comparative analysis of prediction of coronal mass ejections (CME) based on sunspot activities using various machine learning models. In: 2019 International Conference on Robotics, Electrical and Signal Processing Techniques. ICREST, IEEE, pp. 588–591. doi:10.1109/ICREST.2019.8644272.

Marshak, A., Herman, J., Adam, S., Karin, B., Carn, S., Cede, A., Geogdzhayev, I., Huang, D., Huang, L.-K., Knyazikhin, Y., et al., 2018. Earth observations from DSCOVR EPIC instrument. Bull. Am. Meteorol. Soc. 99 (9), 1829–1850. doi:10.1175/BAMS-D-17-0223.1.

Nandy, D., 2021. Progress in solar cycle predictions: Sunspot cycles 24–25 in perspective: Invited review. Sol. Phys. 296 (3), 54. doi:10.1007/s11207-021-01797-2.

National Research Council and Division on Engineering and Physical Sciences and Aeronautics and Space Engineering Board and Space Studies Board and Committee on a Decadal Strategy for Solar and Space Physics (Heliophysics), 2013. Solar and space physics: A science for a technological society. National Academies Press, doi:10.17226/13060.

National Research Council and Division on Engineering and Physical Sciences and Space Studies Board and Committee on the Societal and Economic Impacts of Severe Space Weather Events and A Workshop, 2009. Severe space weather events: Understanding societal and economic impacts: A workshop report. National Academies Press, doi:10.17226/12507.

Pala, Z., Atici, R., 2019. Forecasting sunspot time series using deep learning methods. Sol. Phys. 294 (5), 50. doi:10.1007/s11207-019-1434-6.

Prasad, A., Roy, S., Sarkar, A., Panja, S.C., Patra, S.N., 2023. An improved prediction of solar cycle 25 using deep learning based neural network. Sol. Phys. 298 (3), 50. doi:10.1007/s11207-023-02129-2.

Ramezan, C.A., Warner, T.A., Maxwell, A.E., 2019. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. Remote Sens. 11 (2), 185. doi:10.3390/rs11020185.

Salem, Y.B., Abdelkrim, M.N., 2020. Texture classification of fabric defects using machine learning. Int. J. Electr. Comput. Eng. 10 (4), 4390. doi:10.11591/ijece.v10i4.pp4390-4399.

Saroughi, M., Mirzania, E., Achite, M., Katipoğlu, O.M., Ehteram, M., 2024. Shannon entropy of performance metrics to choose the best novel hybrid algorithm to predict groundwater level (case study: Tabriz plain, Iran). Environ. Monit. Assess. 196 (3), 1–20. doi:10.1007/s10661-024-12357-z.

Schaffer, C., 1993. Selecting a classification method by cross-validation. Mach. Learn. 13, 135–143. doi:10.1007/BF00993106.

Scherrer, P.H., Schou, J., Bush, R., Kosovichev, A., Bogart, R., Hoeksema, J., Liu, Y., Duvall, T., Zhao, J., Title, A., et al., 2012. The helioseismic and magnetic imager (HMI) investigation for the solar dynamics observatory (SDO). Sol. Phys. 275, 207–227. doi:10.1007/s11207-011-9834-2.

Schou, J., Scherrer, P.H., Bush, R.I., Wachter, R., Couvidat, S., Rabello-Soares, M.C., Bogart, R., Hoeksema, J., Liu, Y., Duvall, T., et al., 2012. Design and ground calibration of the helioseismic and magnetic imager (HMI) instrument on the solar dynamics observatory (SDO). Sol. Phys. 275, 229–259. doi:10.1007/s11207-011-9842-2.

Sierra Porta, D., Tarazona-Alvarado, M., 2023. Dataset: Sun dynamics from topological features extraction. Mendeley Data, V1, doi:10.17632/5gh3xbvc92.1.

Singh, A., Bhargawa, A., 2020. Ascendancy of solar variability on terrestrial climate: A review. J. Basic Appl. Sci 16, 105–130. doi:10.29169/1927-5129.2020.16.14.

Solomon, S.C., Liu, H.-L., Marsh, D.R., McInerney, J.M., Qian, L., Vitt, F.M., 2019. Whole atmosphere climate change: Dependence on solar activity. J. Geophys. Res. Space Phys. 124 (5), 3799–3809. doi:10.1029/2019JA026678.

Spencer, D.A., Johnson, L., Long, A.C., 2019. Solar sailing technology challenges. Aerosp. Sci. Technol. 93, 105276. doi:10.1016/j.ast.2019.07.009.

Stone, E.C., Frandsen, A., Mewaldt, R., Christian, E., Margolies, D., Ormes, J., Snow, F., 1998. The advanced composition explorer. Space Sci. Rev. 86, 1–22. doi:10.1023/A:1005082526237.

Tamura, H., Mori, S., Yamawaki, T., 1978. Textural features corresponding to visual perception. IEEE Trans. Syst. Man Cybern. 8 (6), 460–473. doi:10.1109/TSMC.1978.4309999.

Tarazona-Alvarado, M., Sierra-Porta, D., 2023. Dataset for sun dynamics from topological features. Data Brief 51, 109728. doi:10.1016/j.dib.2023.109728.

Tlatov, A.G., 2022. The shape of sunspots and solar activity cycles. Sol. Phys. 297 (8), 110. doi:10.1007/s11207-022-02045-x.

Tougui, I., Jilbab, A., El Mhamdi, J., 2021. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthc. Informat. Res. 27 (3), 189–199. doi:10.4258/hir.2021.27.3.189.

Tsiropoula, G., 2003. Signatures of solar activity variability in meteorological parameters. J. Atmospheric Solar-Terrestrial Phys. 65 (4), 469–482. doi:10.1016/S1364-6826(02)00295-X.

Wang, J., Wiens, J., Lundberg, S., 2021. Shapley flow: A graph-based approach to interpreting model predictions. In: International Conference on Artificial Intelligence and Statistics. PMLR, pp. 721–729, https://ui.adsabs.harvard.edu/link_gateway/2020arXiv201014592W/doi:10.48550/arXiv.2010.14592.

Weerts, H.J., Mueller, A.C., Vanschoren, J., 2020. Importance of tuning hyperparameters of machine learning algorithms. doi:10.48550/arXiv.2007.07588, arXiv preprint arXiv:2007.07588.

Wilson, III, L.B., Brosius, A.L., Gopalswamy, N., Nieves-Chinchilla, T., Szabo, A., Hurley, K., Phan, T., Kasper, J.C., Lugaz, N., Richardson, I.G., et al., 2021. A quarter century of wind spacecraft discoveries. Rev. Geophys. doi:10.1029/2020RG000714.

Wu, C.-M., Chen, Y.-C., 1992. Statistical feature matrix for texture analysis. CVGIP, Graph. Models Image Process. 54 (5), 407–419. doi:10.1016/1049-9652(92)90025-S.

Xiao, Z., 2021. A review of machine learning methods applied in sunspot prediction. In: 2021 International Conference on Networking, Communications and Information Technology. NetCIT, IEEE, pp. 158–161. doi:10.1109/NetCIT54147.2021.00039.

Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 415, 295–316. doi:10.1016/j.neucom.2020.07.061.

Zhang, J., Temmer, M., Gopalswamy, N., Malandraki, O., Nitta, N.V., Patsourakos, S., Shen, F., Vršnak, B., Wang, Y., Webb, D., et al., 2021. Earth-affecting solar transients: A review of progresses in solar cycle 24. Progr. Earth Planetary Sci. 8, 1–102. doi:10.1186/s40645-021-00426-7.