Solar Physics
DOI: 10.1007/●●●●●-●●●-●●●-●●●●-●

Analysis of solar minima using machine learning techniques

ArnavRanjekar · YaduKrishnan · Tarun K. Jha

Abstract Understanding solar activity and its impact is essential for space weather prediction and technological systems. This study employs machine learning (ML) techniques to analyze and forecast sunspot activity, focusing on hemispheric sunspot numbers and 10.7 cm solar flux. The pre-processing includes handling missing values, aggregating daily counts into monthly averages, and normalization. Using advanced ML models, particularly an RNN-LSTM hybrid, the study achieved significant predictive accuracy, with RMSE of 21.37 and MAE of 16.37, corresponding to 13.48% error relative to average sunspot numbers. Fourier analysis confirmed the Schwabe cycle (11 years) as the dominant signal, and anomaly detection using Isolation Forest uncovered significant outliers in sunspot data. The results validate the model's capability to capture solar cycles' cyclic nature, offering a robust method for solar activity prediction and anomaly detection.

Keywords: Sunspot cycles, Solar activity, Machine learning, RNN-LSTM hybrid, Schwabe cycle, Sunspot prediction, Time-series analysis, 10.7 cm solar flux, Anomaly detection, Fourier analysis, Hemispheric sunspot numbers, Space weather, Solar forecasting.

1. Introduction

Studying solar activity is crucial for several reasons, particularly in the context of space weather and its impact on various systems. Understanding the sun's behavior can help us anticipate solar flares and coronal mass ejections, which can have significant effects on space probes, satellite operations, and even power

✉ A.Ranjekar
arnav@email.com
Y.Krishnan
yadu@email.com
T.Jha
tkj@email.com

grids on Earth. Here we mainly see about the magnetic activity of the sun and thereby photosphere via sunpots and corona with the 10.7cm flux. Solar flux at 10.7cm radio emission-Radio waves are mainly emitted by the corona. With increase in activity this emission also becomes intense and during minima this is low(Okoh and Okoro,2020) Increased solar activity can disrupt communication systems and navigation technologies, leading to potential hazards for astronauts and spacecraft. By gaining insight into solar cycles and their influence on Earth's climate and atmospheric conditions, we can enhance our predictive capabilities regarding space weather events ensuring the safety and effectiveness of space exploration and daily life on Earth. Studying solar minima is crucial for understanding the Sun's influence on Earth, the solar system, and space weather.
Solar minima represent the period of least solar activity within the 11-year solar cycle, providing a baseline to measure variations and anomalies during solar maxima. The Sun's magnetic field weakens and the heliosphere contracts, allowing more cosmic rays to penetrate the inner solar system which presents increased risks for astronauts, satellites, and high-altitude flights. Reduced solar activity can lead to cooling and contraction of the upper atmosphere. This affects satellite orbits, as decreased atmospheric drag might prolong satellite lifetimes but could also raise the risk of space debris collisions. Solar minima are associated with lower solar irradiance, potentially impacting Earth's climate. Historical prolonged solar minima, like the Maunder Minimum during the 17th century, have been linked to cooler periods such as the "Little Ice Age." Researching solar minima helps in assessing total Solar Irradiance, vital for understanding solar contributions to climate change, helping to distinguish them from human-induced effects. The increased influx of cosmic rays and reduced solar wind shielding during solar minima can provide valuable insights into how solar activity might affect the habitability of planets, particularly those with thin or absent atmospheres. By studying solar minima, we gain a more comprehensive understanding of solar dynamics and their broader implications across scientific, technological, and environmental fields.

2. Data

Data Preprocessing The data preprocessing on the daily sunspot number csv was done in 3 stages

2.1. Validation of data

1 . Handling of missing values The dataset contained sunspot counts in which invalid values of -1 represented missing or unreliable data points. The values were replaced with NaN to ensure accurate analysis. This step was critical to prevent distortions in statistical calculations and visualizations. The active decision to replace -1 with NaN was taken, instead of interpolating as that would introduce artificial values that could disrupt the natural cycllic patterns in susnpot data, especially in prolonged periods of missing values. Interpolation would assume a continuity or linearity in trends which isn't true for highly variable and cyclic

datasets like sunspot numbers. It is relatively easy to implement models that can handle NaN values reliably, which negates the need for interpolation as well. 2. Aggregation to Monthly Averages Daily sunspot counts were aggregated into monthly averages using a resampling technique. This conversion was necessary so as to avoid noise and focus on long-term trends such as the Schwabe cycle. 3. Feature Scaling StandardScaler was applied to normalize the Sunspot Number. Unlike MinMaxScaler(which was initially used) , StandardScaler is not sensitive to outliers and ensures proper normalization which is important for model performance and ensuring better convergence during training. Temporal Data Indexing Datatime indices were created for the dataset to facilitate time-series operations which ensured effective handling of temporal dependencies in the data.

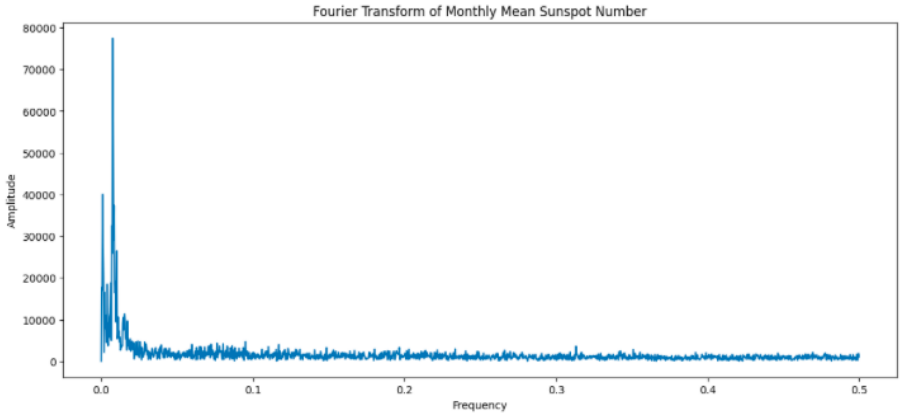

Figure 1. Fourier transform is used to decompose the time series signal into its constituent frequencies , so that the dominant periodic components of the solar activity could be analyzed. The graph plotted (horizontal axis is frequency present in the sunspot data, measured in cycles per month and vertical axis is amplitude of each frequency component) shows a dominant low -frequency component (near 0 on the x-axis) which indicates the present of a long -period cycle corresponding to the well-known Schwabe cycle(0.0076 cycles per month). As the frequency increases, the amplitude decreases suggesting that high-frequency noise or rapid fluctuations contribute less to the overall variability in sunspot activity. Through the plot we were able to show that sunspot data predominantly exhibits low=frequency , long term cyclical behaviour, aligning with theoretical expectations of solar Activity cycles.

2.2. Sources

SILSO, World Data Center - Sunspot Number and Long-term Solar Observations, Royal Observatory of Belgium, on-line Sunspot Number catalogue: http://www.sidc.be/SILSO/
The results presented in this document rely on data collected by the Solar Radio Monitoring Program (https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-en.php) operated by the National Research Council and Natural Resources Canada. These data are available at https://www.spaceweather.gc.ca/forecast-prevision/solar-solaire/solarflux/sx-5-en.php. These data were accessed via the LASP Interactive Solar Irradiance Datacenter (LISIRD (https://lasp.colorado.edu/lisird/).

3. Working Methodology

In a comprehensive literature review, a comparative analysis was performed on various machine learning models used to predict sunspot data, focusing on their performance metrics, specifically Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The study examined several models, including K-Nearest Neighbors (KNN), Random Forest, LightGBM (LGBM) and XGBoost, to assess their predictive capabilities. Additionally, the review highlighted the application of deep neural networks and recurrent neural networks (RNN) in sunspot predictions, emphasizing the potential benefits of these advanced techniques in capturing the underlying patterns in temporal data. Such analyses are crucial for understanding the efficacy of different methodologies in astrophysical forecasting. An addition to our paper will be that we are using hemispheric sunspot numbers as well. Other relevant data has been used as well, including the 10.7 cm solar flux, plage area, to check their correlation with sunspot activity. Additionally, we will explore how these factors influence solar activity patterns and assess their implications for understanding solar cycles more comprehensively. The integration of these diverse data sets will enhance the predictive capabilities of our model.

3.1. Justification for method used

Using machine learning, particularly recurrent neural networks (RNNs), for analyzing sunspot data is advantageous because of the unique characteristics of the data and the analytical needs of solar studies. Sunspot data consists of observations over time, making it a time-series dataset. RNNs are specifically designed for analyzing sequential data because they can retain and use information from previous time steps, allowing them to capture temporal dependencies and patterns. Sunspot cycles are governed by the Sun's magnetic dynamo, which involves highly nonlinear and complex physical processes. Traditional statistical models may struggle to capture these relationships, whereas machine learning models, especially neural networks, excel in modeling such complexities.RNNs can identify hidden patterns and periodicities in sunspot data. Historical sunspot data often contains missing values or noise. Machine learning models can be trained to handle these imperfections better than traditional models by learning robust features and using techniques like data imputation. Beyond just sunspot counts, solar data may include associated variables like solar irradiance, magnetic field strength, and geomagnetic indices. Machine learning models, including RNNs, can process multivariate time-series data effectively to uncover deeper insights.

3.2. ML model

This study explores sunspot data from a single daily sunspot data set which was aggregated into monthly averages, employing statistical analysis , signal processing and machine learning techniques to study the patterns , detect the anomalies and predict future trends. The light weight model, designed to operate

using data from only one source, achieved a RMSE of 21.37 and a Mean Absolute Error of 16.37 corresponding to 13.48 % of the average sunspot count. Despite its simplicity the model delivers good performance capturing the cyclic nature of sunspot activity effectively. The findings confirm the dominant presence of the Schwabe cycle and demonstrate the feasibility of accurate sunspot activity predictions.

Sunspot Prediction: Builds an RNN-LSTM hybrid model to forecast monthly sunspot numbers.

3.2.1. Anomaly Detection Using Isolation Forest

Isolation Forest, with a contamination level of 5%, identified outliers. Normal data and anomalies were visualized over time to uncover patterns. Daily sunspot counts with values of -1 were treated as missing and excluded. The dataset was restructured with datetime indices for temporal analysis. Fourier analysis confirmed the Schwabe cycle (11 years) as the dominant periodic signal. The frequency and amplitude metrics underscored the strong influence of this cycle. Key findings: Anomalies: Significant outliers corresponded to unusually high or low sunspot counts. Rolling Means and KDE Analysis: Highlighted smoothing trends and density distributions of anomalies vs. normal data.

3.2.2. RNN architecture

Monthly mean sunspot numbers were scaled using StandardScaler. Sequence data with a 36-month lookback was generated for the RNN-LSTM model. An enhanced RNN - LSTM hybrid model was used with the following layers with the ability to capture long-term dependencies. Dropout addressed the high variability in sunspot activity and reduces overfitting while the batch normalization ensures model stability while accelarating the training. The rnn and lstm combination was chosed to balance computational effeciency SimpleRNN layer - captures short-term dependencies using a recurrent structure with low computational overhead. LSTM layers - Layer 1(128 units): Extracts long-term temporal patterns with a wide receptive field) Layer 2(64 units) : Refines extracted features, focusing on complex dependencies . Dropout and Batch Normalization : Added after each major layer to prevent overfitting and stabilized learning. Dense Layers : Intermediate dense layer (32 units): Convertes extracted features into a lower-dimensional representation Output Layer(1 unit): Predicts the next month's sunspot count. Step 3 : Training Loss function : Huber Loss was selected as it can handle outliers in the sunspot data. Optimization : Adam Optimizer with a learning rate scheduler adjusted the learning rate dynamically based on validation performance. Early stopping : monitored the validation loss , and would halt if there was no improvement in 10 consecutive Epochs. This ensured that the model wouldn't overfit. The model was trained on data prior to 2019 and validated on subsequent data.

3.2.3. Results:

Validation RMSE: 21.37 Predicted trends closely followed actual observations, capturing the overall cyclic nature.

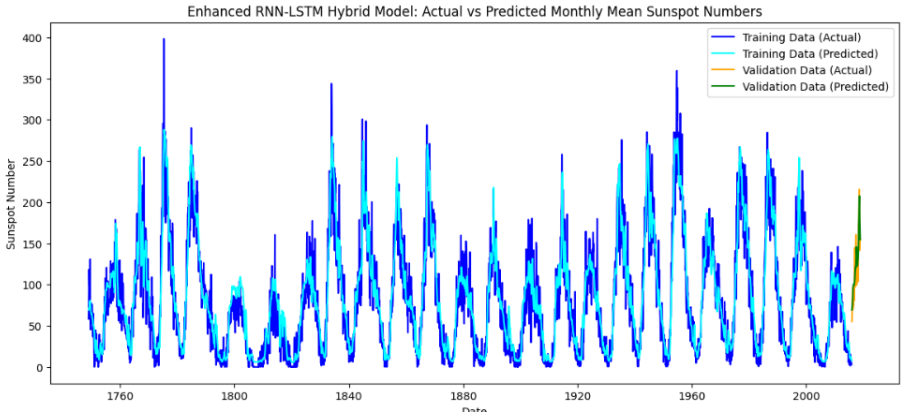

Figure 2. From the graph we are able to visualize the accuracy of the model. The predicted sunspot numbers closely follow the actual values both in the training and validation datasets showing strong agreement with historical patterns. The model managed to successfully capture the cyclic nature of sunspot activity , including the peaks and troughs characteristic of the solar cycle.

4. Results & Discussions

The integration of machine learning and signal processing allowed for robust anomaly detection and trend forecasting. Limitations include: Sensitivity to extreme values in training. Simplifications in modeling solar activity complexities. We were able to achieve an absolute rmse of 21.37 which translates to a percentage rmse of 17.67 percent which highlights the effectiveness of the model in capturing the cyclic nature of the sunspot activity. An MAE of 16.37 was calculated which suggests the errors are distributed without extreme outliers. This translated to a 13.48 percent deviation which implies the model's predictions are closely aligned with the observed values and the model was able to capture both trends and cycles effectively.This result is particularly notable given that the analysis and prediction was conducted on a single daily sunspot dataset. Despite the inherent challenges posed by working with a single data source, the model demonstrated a strong ability to generalize and predict sunspot numbers accurately.

5. Conclusion

This study demonstrates the effectiveness of integrating machine learning and signal processing techniques in understanding and predicting solar activity. By leveraging an RNN-LSTM hybrid model, it successfully captured the cyclic nature of sunspot activity and provided accurate predictions even with a single data source. The study highlighted the Schwabe cycle's dominance and identified significant anomalies, underscoring the utility of ML for astrophysical forecasting. Despite challenges such as data limitations and solar activity complexities, the model exhibited strong generalization and predictive performance. Future work

could enhance these findings by incorporating additional datasets and exploring the impact of other solar parameters on sunspot activity, further advancing the predictive modeling of solar cycles.

6. References

1. Solar Physics. Data-Driven Forecasting of Sunspot Cycles: Pros and Cons of a Machine-Learning-Based Approach, 2024. DOI: 10.1007/s11207-024-02270-6.

2. Scientific Reports. Probabilistic Sunspot Predictions with a Gated Recurrent Units-Based Model, 2024. DOI: 10.1038/s41598-024-63878-z.

3. ResearchGate. Prediction of Yearly Mean Sunspot Number using Machine Learning Methods, 2023. Accessible at: https://www.researchgate.net/publication/382911222.

4. Solar Physics. Hemispheric Sunspot Number Prediction for Solar Cycles 25 and 26 Using Machine Learning and Time Series Spectral Analysis, 2024. DOI: 10.1007/s11207-024-02363-2.

5. Electronics. Prediction of Sunspot Number with Hybrid Model Based on 1D-CNN, 2023. DOI: 10.3390/electronics13142804.

6. Frontiers in Physics. A Hybrid Model for Forecasting Sunspots Time Series Based on Variational Mode Decomposition, 2018. DOI: 10.3389/fphys.2018.01326.

7. arXiv. A Comparative Study of Non-Deep Learning, Deep Learning, and Ensemble Learning Methods for Sunspot Number Prediction, 2022. arXiv:2203.05757.

8. arXiv. Deep Learning Reconstruction of Sunspot Vector Magnetic Fields for Forecasting Solar Storms, 2022. arXiv:2209.09944.

9. arXiv. Forecasting Solar Cycle 25 using Deep Neural Networks, 2020. arXiv:2005.12406.

10. arXiv. A Model-Free, Data-Based Forecast for Sunspot Cycle 25, 2020. arXiv:2005.12166.