Predictive Modeling of Solar Activity: Analysing Daily Sunspot Data (1818–2025) Using Machine Learning and Time Series Forecasting

Student’s Name

Date

**Table of Contents**

[Introduction 3](#_Toc192604529)

[Overview and Purpose of the Project 3](#_Toc192604530)

[Research Question, Aims, and Objectives 4](#_Toc192604531)

[Research Question 4](#_Toc192604532)

[Aim 4](#_Toc192604533)

[Objectives 4](#_Toc192604534)

[Background 4](#_Toc192604535)

[Selection Criteria for Literature Review 4](#_Toc192604536)

[Paper 1: “A Comparative Study of Non-Deep Learning, Deep Learning, and Ensemble Learning Methods for Sunspot Number Prediction” 5](#_Toc192604537)

[Paper 2: “Synergistic Sunspot Forecasting: A Fusion of Time Series Analysis and Machine Learning” 6](#_Toc192604538)

[Paper 3: “Solar Cycle Prediction Using a Long Short-Term Memory Deep Learning Model” 6](#_Toc192604539)

[Paper 4: “Sunspots Time-Series Prediction Based on Complementary Ensemble Empirical Mode Decomposition and Wavelet Neural Network” 7](#_Toc192604540)

[Summary of Literature Review 8](#_Toc192604541)

[Methodology 9](#_Toc192604542)

[References 11](#_Toc192604543)

# Introduction

## Overview and Purpose of the Project

Space weather affects satellite communications, power grids, and other technological systems on Earth. Solar activity, primarily measured by sunspot numbers, has a significant role in determining the levels of space weather. Accurate solar activity forecasting is essential for these systems' possible adverse effects (Horat, Klerings and Lerch, 2025). This study aims to use machine learning techniques and time series forecasting approaches to develop predictive models for the daily sunspot numbers in 1818 - 2019. Solar activity is known to have long been a significant factor in the space weather conditions. Sunspots are temporary phenomena in the photosphere of the Sun that are indicators of solar magnetic activity. However, the number of sunspots varies in an approximately 11-year solar cycle (Ledmaoui *et al.*, 2023). When sunspots are abnormally high during periods of high solar activity, they see increased sun radiation and sun flares that mess with satellite communications, scramble power grids, and place astronauts and spacecraft in danger. The prediction of solar activity is significant in preserving the technological infrastructure and guaranteeing the safety of space missions.

They have monitored and recorded solar activity for centuries, and sunspot numbers are one of the oldest and longest-running records of solar activity (Doumèche *et al.*, 2025). This project uses a database from 1818 to 2019, which contributes with a long-time range perspective about solar activity. Such an extensive dataset ensures that one can develop robust predictive models capable of describing patterns and trends of solar activity over time. Several methods were utilised in the industry to forecast solar activity. The solution is generally addressed by traditional approaches that rely on statistical models and empirical relationships, which are constructed with historical data (Alvarado, Herrera Acevedo and Porta, 2024). Nevertheless, since these methods cannot reproduce the dynamics of complex and non-linear solar activity, additional methods are required to capture the effects of the significant changes in solar activity on each of these variables. The view of applying machine learning and advanced computational techniques to improve the accuracy and reliability of solar activity forecasts has grown with the advent of machine learning and advanced computational methods (Horat, Klerings and Lerch, 2025).

Complex time series data and the patterns embedded within, often invisible to traditional methods, can be handled by machine learning algorithms such as neural networks and ensemble methods. Based on historical data, time series forecasting techniques, including autoregressive integrated moving average (ARIMA) models, and newer deep learning approaches like Long Short-Term Memory (LSTM) networks allow the future values to be predicted using the historical data (Li and Law, 2024). This project combines advanced techniques to create predictive models to accurately forecast daily sunspot numbers, thereby enabling us to predict and minimise the effects of solar activity on Earth's technological systems.

## Research Question, Aims, and Objectives

### Research Question

* Can machine learning enhance predictive accuracy beyond traditional statistical models?

### Aim

* The study aims to analyse solar activity by performing predictive modelling using machine learning and time series forecasting.

### Objectives

* Preprocessing and analysing historical sunspot data from 1818 to 2019.
* To evaluate and compare the performance of various machine learning and time series forecasting models in predicting sunspot numbers.
* Identify the most effective model(s) for accurate daily sunspot prediction.
* To provide insights into the patterns and cycles of solar activity based on model predictions.

# Background

Sunspot regions of the solar surface characterised by extreme magnetic activity have been an essential focus of solar physics because they affect space weather and terrestrial technologies (Ishii *et al.*, 2024). Methods of forecasting traditional means of sunspot prediction have transitioned from statistical due to more advanced machine learning techniques. This review covers the literature that uses machine learning or time series forecasting methods for sunspot prediction and provides a comparative study of various predictive models while analyzing how to preprocess and endure long-term solar activity data.

## Selection Criteria for Literature Review

Different ways have been used to tackle forecasting challenges within sunspot prediction, generally linear and non-linear modelling. The time series data of the sunspot numbers shows some unique properties: uncertainty, volatility and cyclicity. Non-linear modelling techniques are found more suitable for sunspot number forecasting. Classical statistical methods and neural network techniques have been widely employed for non-linear modelling studies on sunspot number prediction (Alvarado, Herrera Acevedo and Porta, 2024). Additionally, multiple technologies have been combined in one model to solve the time series forecasting problems. Ishii *et al.* (2024) analysed various forecasting time series models, specifically the autoregressive integrated moving average (ARIMA) approach to this problem along with the dynamic neural network (DNN) approach and, out of all, they found the latter to provide superior forecasting time series accuracy.

Another combined model integrates ARIMA and a support vector machine (SVM) to predict monthly and yearly sunspot numbers. Ishii et al. (2024) presented a modified artificial neural network (ANN)-ARIMA forecasting model using bootstrap methods to increase the point accuracy and speed of the sunspot time series forecasting. Li and Law (2024) solved the non-linear part of the sunspot time series using the multi-layer perceptron model (MLP). The method provided by Pala et al. is a combined approach that integrates neural network autoregression and neural network architecture of long short-term memory (LSTM) to process the time series data of sunspot numbers. They suggest using hybrid methods to improve the overall process and performance of the hybrid model for better results (Ledmaoui *et al.*, 2023).

### Paper 1: “A Comparative Study of Non-Deep Learning, Deep Learning, and Ensemble Learning Methods for Sunspot Number Prediction”

According to Dang *et al.* (2022), this study offers a complete analysis of non-deep learning, deep learning, and ensemble methods for the prediction of sunspot numbers that provide valuable insights into the achievement of predicting solar activities using different models. An innovative ensemble model, the XGBoost-DL lifts deep learning models using XGBoost to produce a two-level non-linear combination. The RMSE and MAE of all other models tested are below 30, which is better than the RMSE and MAE of 25.70 and 19.82, respectively, of the proposed model (Dang *et al.*, 2022). One of many critical foundational aspects of this research is that it had to rely on historical sunspot data to train and test to train and test the models. Although this offers a solid basis for assessing the model's performance over time series data, it also points out potential data quality and availability limitations. Monthly sunspot numbers are the primary focus of the study, which may not incorporate all variability in the data, and this could be addressed by taking the approach to finer temporal resolutions. For Solar Cycles 25 and 26, the XGBoost-DL model further predicts their peak sunspot numbers at 133.47 and 164.62 (Dang *et al.*, 2022). They are similar but slightly later than the predictions made by NASA. As a whole, this work provides a strong framework for predictive modelling in solar physics. Still, as discussed, more granular temporal scales would improve predictions under different situations.

### Paper 2: “Synergistic Sunspot Forecasting: A Fusion of Time Series Analysis and Machine Learning”

Menghui Chen et al. (2024), in the study "Synergistic Sunspot Forecasting: A Fusion of Time Series Analysis and Machine Learning", has provided a broad methodology for predicting the day-to-day sunspot data by joining the non-linear time series analysis with machine learning algorithms. For the time series analysis, the authors use the Hurst exponent as well as the fast Fourier transform, together with Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) models for prediction (Chen *et al.*, 2024). Based on these data, the research leverages daily sunspot data over many solar cycles, thus providing a source for evaluating model performance. Traditional time series methods are combined with the most powerful machine learning algorithms to characterise the data from linear and non-linear perspectives. The results show that the higher the R squared and the mean absolute error (MAE), the better the outcomes are compared to standalone models. This work fits the design of other projects to increase forecasting accuracy in solar physics. This means a critical analysis still needs further exploration of preprocessing techniques to obtain better performance and optimise the model. These combined models can then be used in future work to study how the predictive power of these combined models changes with different preprocessing strategies (Chen *et al.*, 2024).

### Paper 3: “Solar Cycle Prediction Using a Long Short-Term Memory Deep Learning Model”

The drawbacks of the other proposed models are tackled in the current proposed study of "Solar Cycle Prediction Using a Long Short Term Memory Deep Learning Model," which suggests an optimal LSTM model for the prediction of the smoothed monthly sunspot number (SSN) (Wang, Li and Guo, 2021). It uses the historical SSN data available on different solar cycles to optimise the LSTM architecture parameters, hidden nodes, and batch size. The one-step and multi-step models are also used to predict Solar Cycles 22 to 24, with the maximum RMSE achieved by the one-step model as 6.12. However, the study results suggest that LSTM models can capture the temporal dependency for the sunspot data and are good enough for temporal dependency prediction. However, this may not be an obvious case for the daily data as the daily data contain a higher noise level than the smoothed monthly data.

Furthermore, the emergence of Solar Cycle 25 is forecasted to reach a peak amplitude of approximately 114.3 around 2023, and this is different from other predictions, which implies the need for validation with updated datasets (Wang, Li, and Guo, 2021). The work has also been subjected to a critical analysis, which shows that although this work shows a good demonstration that LSTMs can be used to predict solar cycles, more work is required to make these models feasible for real-time or daily forecasting. This prediction was also in contrast to other predictions; hence it emphasises that an ongoing model validation and comparison with emerging forecasts from other methodologies like hybrid CNN-LSTM or LSTM-WGAN models is important (Wang, Li and Guo, 2021).

### Paper 4: “Sunspots Time-Series Prediction Based on Complementary Ensemble Empirical Mode Decomposition and Wavelet Neural Network”

The scientists from China made a proposal in the study 'Sunspots Time Series Prediction Based on Complementary Ensemble Empirical Mode Decomposition and Wavelet Neural Network' Li and Wang (2017), where sunspot activity is predicted through Complementary Ensemble Empirical Mode Decomposition (CEEMD) and Wavelet Neural Networks (WNN) (Li and Wang, 2017). Since sunspot data is non-linear and nonstationary, extra effort must be made to improve prediction accuracy. The hybrid model has been devised to cater to this. The authors then decompose the sunspot time series into intrinsic mode functions (IMFs) and use those IMFs as inputs to the WNN. This method is capable of capturing the more intricate pattern in the data body (Li and Wang, 2017). It is found that the performance of the CEEMD-WNN model is superior when compared with the other methods for an RMSE and an MAE of 12.64374 and 1.58413, respectively. This study shows a way to overcome the nonlinearity in sunspot data to predict daily, which is the relevant task. Even so, a critical analysis reveals that there could be a constraint of computational complexity which delays real-time daily prediction due to higher processing requirements. The last is to advance this hybrid model further to increase efficiency without losing predictive power. (Li and Wang, 2017).

## Summary of Literature Review

Studies use many ways to predict solar behaviour because each research team tackles specific problems when forecasting sunspot data over time. According to Dang *et al.* (2022), their XGBoost-DL ensemble model shows outstanding performance in predicting sunspot numbers by achieving lower error rates at 25.70 RMS and 19.82 MAE (Alvarado, Herrera Acevedo, and Porta, 2024). Although their model accurately predicts Solar Cycle events, it needs daily data instead of monthly aggregates because short-term solar variations matter for operational space weather monitoring. Chen et al. (2024) framework linking Hurst exponent analysis and neural networks with Fast Fourier Transform helps predict solar behaviour better than independent models (Doumèche *et al.,* 2025). They use multi-solar-cycle high-frequency data but suggest ways to reduce noise at the start of their methods (Ishii *et al.,* 2024). Wang *et al.* (2021) discovered that LSM performances decline in noisy daily databases despite their optimised LSTM structure providing an RMSE result of 6.12 on smoothed monthly data (Alvarado, Herrera Acevedo, and Porta, 2024). The Solar Cycle 25 peak prediction result from Wang et al. deviates significantly from NASA consensus values because their model reacts strongly to the quality of input data and methods used to convert time series into aggregated values. Li and Wang (2017) show how combining CEEMD-WNN improves forecasting accuracy by first splitting sunspot records into intrinsic mode functions before neural network processing leads to RMSE 12.64 on daily projections (Wang, Li, and Guo, 2021). Their two-step forecasting method would create challenges because of its high processing needs in current operational systems. These investigations show how academic research now develops mixed approaches that connect signal processing tools with machine learning yet deal with performance versus complexity concerns (Alvarado, Herrera Acevedo, and Porta, 2024). The literature review defines basic analysis methods and finds essential stay-in-reach areas for noise reduction algorithms, faster processing development, and model testing standards.

# Methodology

I started my solar activity prediction project by adding necessary libraries to manage data movement and visualisation while also applying time series and machine learning methods. For my data manipulation, I picked pandas and numpy then used matplotlib and seaborn to make visualisations and statsmodels packages plus Prophet and other specific libraries to develop forecasts. ARIMA, Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) and Vector autoregression (VAR) techniques also supported this process. I designed my time series project with TensorFlow deep learning in mind and selected Trigonometric seasonality, Box-Cox transformation, ARIMA errors, Trend and Seasonal (TBATS) and pmdarima to provide more robust time series decomposition while mostly relying on the LSTM approach. I uploaded daily sunspot historical data from 1818 to 2019 into the system through the “sunspot\_data.csv” file. To begin my analysis, I studied the dataset structure and printed its initial rows while evaluating distribution patterns, total data range, and unusual value behaviour. I tested to see if the dataset contained empty entries because this affects the quality of model learning. The analysis demands precise handling of time-related data so I transformed the date column to datetime format sorted the data chronologically and made the date column the primary reference. Transforming the data in this way made it easier to perform time-based evaluations and create visual representations.

My initial data review consists of plotting sunspot numbers against time to find patterns and irregularities that naturally occur due to the 11-year solar cycle. At first, I wanted to divide the dataset between records before and after January 1, 2019, based on date. To build more resistant testing and training methods I split my data into 80 per cent training and 20 per cent testing. The MinMaxScaler from sci-kit-learn enables the normalisation of sunspot numbers when dealing with their non-stationary trend. An LSTM network needs normalisation to train efficiently since it regulates the numerical values and makes their scales uniform. I implemented a custom sequence generator function called create\_sequences which then transformed the normalised time series into sequences. I chose the sequence length to be 12 for each prediction, which is the preceding 12 observations. However, it was crucial for capturing the temporal dependencies in the data and the temporal dynamics that were present in the data.

After reshaping to the format [samples, time steps, features] required by the LSTM, these sequences were passed into the LSTM. To construct the neural network, I constructed a sequential model, which consists of two LSTM layers, with 50 units in each. To counter overfitting, I incorporated dropout layers brought in with a dropout rate of 0.2 between LSTM layers, randomly deactivating a total of 20% of neurons during training. The final regression output from the given network consists of a Dense layer at the end. In particular, I was using MSE here as a regression loss function, because MSE is highly sensitive to large errors, and it's frequently used in regression tasks and is accepted enough. The model was trained for 50 epochs with a batch size of 32 and with 10% of the training data for validation to monitor model performance to prevent overfitting. After training, I then used the test set to generate forecasts using the LSTM model and then ran the inverse transformation of the scaler on the prediction to put the values back into the sunspot number scale. I plotted the actual vs forecasted sunspot numbers to verify that the model is performing as expected, while each step (from data preprocessing to evaluation of the model) has been optimally optimised to maximise the prediction accuracy of the daily solar activity.

# References

Alvarado, M.T., Herrera Acevedo, D.D. and Porta, D.S. (2024) ‘Toward Solar Activity Understanding: Multifaceted Insights from SOHO Solar Images and Spectral Analysis for Machine Learning Applications’, in *2024 XVIII National Meeting on Optics and the IX Andean and Caribbean Conference on Optics and its Applications (ENO-CANCOA)*. *2024 XVIII National Meeting on Optics and the IX Andean and Caribbean Conference on Optics and its Applications (ENO-CANCOA)*, pp. 1–6. Available at: <https://doi.org/10.1109/ENO-CANCOA61307.2024.10751088>.

Chen, M., Kumarasamy, S., Srinivasan, S. and Popov, V. (2024) ‘Synergistic sunspot forecasting: a fusion of time series analysis and machine learning’, *Pramana*, 99(1), p. 9. Available at: <https://doi.org/10.1007/s12043-024-02861-9>.

Dang, Y., Chen, Z., Li, H. and Shu, H. (2022) ‘A Comparative Study of non-deep Learning, Deep Learning, and Ensemble Learning Methods for Sunspot Number Prediction’, *Applied Artificial Intelligence*, 36(1), p. 2074129. Available at: <https://doi.org/10.1080/08839514.2022.2074129>.

Doumèche, N., Bach, F., Bedek, É., Biau, G., Boyer, C. and Goude, Y. (2025) ‘Forecasting time series with constraints’. arXiv. Available at: <https://doi.org/10.48550/arXiv.2502.10485>.

Horat, N., Klerings, S. and Lerch, S. (2025) ‘Improving Model Chain Approaches for Probabilistic Solar Energy Forecasting through Post-processing and Machine Learning’, *Advances in Atmospheric Sciences*, 42(2), pp. 297–312. Available at: <https://doi.org/10.1007/s00376-024-4219-2>.

Ishii, M., Berdermann, J., Forte, B., Hapgood, M., Bisi, M.M. and Romano, V. (2024) ‘Space weather impact on radio communication and navigation’, *Advances in Space Research* [Preprint]. Available at: <https://doi.org/10.1016/j.asr.2024.01.043>.

Ledmaoui, Y., El Maghraoui, A., El Aroussi, M., Saadane, R., Chebak, A. and Chehri, A. (2023) ‘Forecasting solar energy production: A comparative study of machine learning algorithms’, *Energy Reports*, 10, pp. 1004–1012. Available at: <https://doi.org/10.1016/j.egyr.2023.07.042>.

Li, G. and Wang, S. (2017) ‘Sunspots Time‐Series Prediction Based on Complementary Ensemble Empirical Mode Decomposition and Wavelet Neural Network’, *Mathematical Problems in Engineering*. Edited by T. Kapitaniak, 2017(1), p. 3513980. Available at: <https://doi.org/10.1155/2017/3513980>.

Li, W. and Law, K.L.E. (2024) ‘Deep Learning Models for Time Series Forecasting: A Review’, *IEEE Access*, 12, pp. 92306–92327. Available at: <https://doi.org/10.1109/ACCESS.2024.3422528>.

Wang, Q.-J., Li, J.-C. and Guo, L.-Q. (2021) ‘Solar cycle prediction using a long short-term memory deep learning model’, *Research in Astronomy and Astrophysics*, 21(1), p. 012.