HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 1 (2023)

A Hybrid CNN-LSTM Model for Predicting Solar Cycle 25

Alice Hu[1] and Antonio Rodriguez[#]

[1]West Windsor-Plainsboro High School North, Plainsboro Township, NJ, USA
[#]Advisor

ABSTRACT

The solar cycle is linked to the number of sunspots and follows the fluctuations of the Sun's magnetic field. It can have powerful global impacts on the Earth. Thus, predicting the timing and amplitude of the peak of the incoming solar cycle 25 is of great importance. This study uses a hybrid deep learning convolutional neural network (CNN) - long short-term memory (LSTM) model and the observed 13-month smoothed sunspot numbers to predict Solar Cycle 25. Here it is shown for the first time that the MinMax normalization method substantially reduces the error of the CNN-LSTM model's solar cycle predictions compared to the Standard Deviation normalization method. The results also suggest that it is best to use four historical solar cycles to predict the future solar cycle. The predicted Solar Cycle 25 has a 13-month smoothed peak amplitude similar to that of Solar Cycle 24. The predicted Solar Cycle 25 peak spans a relatively long period of time between approximately August 2023 and July 2024.

Introduction

The sunspot induced by the Sun's magnetic field goes through a solar cycle of approximately 11 years. The formation and cycles of the magnetic fields and sunspots on the Sun are common sources of curiosity for solar physicists. In addition to causing sunspots, the Sun's magnetic field also controls the motion of its corona, the outer layers above the visible surface of the Sun, which produces solar wind, solar flares, and coronal mass ejections (CMEs) due to the twisting of magnetic field lines, and releases large amounts of energy. The particles from solar wind and other solar events often interact with the Earth's magnetosphere, producing geomagnetic storms. The Earth's radiation belts contain the highest energy plasma, and can be powerful enough to pose a threat to satellites. The state of the upper layer of the thermosphere, the ionosphere, is important because it scatters and absorbs radio waves. The conditions of the ionosphere affect technology like Global Positioning System (GPS) and communication systems. Solar activity is responsible for the aurora, which increases the wind strength and temperature in the thermosphere. Solar wind also makes up the heliosphere, which protects the planets from cosmic rays (National Research Council, 2013). The escape and ionization of atoms in the Earth's atmosphere are directly related to solar UV radiation and the strength and magnetism of the solar wind. It is theorized that large enough solar flares and their resulting geomagnetic storms could destroy parts of the ozone layer across the globe and eliminate certain diatomic species (Moore, 2020). The interactions between solar storms and the Earth's magnetic field affect the functionality of electrical conductors and induce large currents that damage modern technology like power distribution systems. There is also evidence that solar activity affects the Earth's climate, causing temperatures to rise and fall. The Sun and its output will be instrumental to determining the future of planet Earth.

Solar cycle predictions will help us prepare ahead of time for future solar activity. Yet, despite the solar cycle's significance, very little is known about the solar cycle. It remains challenging to make accurate predictions of the timing and amplitude of the solar cycle. Increases in sunspots correspond to more x-ray and

ultraviolet emissions from the Sun and thus more geomagnetic storms affecting the orbital path of satellites. Major solar flares cause widespread damage to satellites and electrical grids during the peak of the solar cycle. Now more than ever, the world's internet infrastructure is very vulnerable to outages in the event of large-scale solar events like CMEs, especially undersea cables. The impact could last several months and be global in scale. And, electrical failures are costly to repair. According to some estimates, the 2012 CME had cost the US alone $2.6 trillion (Jyothi, 2021). Being able to predict the solar cycle ahead of time allows us to mitigate the effects of major solar flares and CMEs. For example, since increased solar activity increases atmospheric resistance and negatively affects satellites, knowing when the solar maximum occurs gives us a timeframe of when satellites should not be sent into space and when adjustments are needed.

There have been two distinct ways to approach solar cycle/sunspot predictions: the traditional statistical and/or theoretical methods, and the deep learning methods. Among the current non-deep learning methods used to predict the solar cycle, many are sensitive to input data information, such as the choice of the solar cycle minimum and how the data is smoothed. The geomagnetic precursors occur much later than the solar cycle minimum, limiting its prediction capability (Hathaway, 2015). With improved observation and analysis capabilities, previous studies can find out where the releases of solar energy are likely to occur, but still cannot consistently pinpoint when and how large it will be. This is concerning since the timing and magnitude of a solar cycle peak are the most significant factors in determining and mitigating the effects of solar activity on us. The currently predicted Solar Cycle 25 sunspot numbers released from the international panel co-chaired by NOAA/NASA are lower than those observed, leading to the urgent need for improving sunspot prediction methods. Deep learning models are more effective at predicting the nonlinear solar cycles than the traditional statistical models, such as the linear regression model. As for the deep learning approach, recurrent neural networks (RNN) are especially appropriate because the sunspot cycle has an approximate period of 11 years and includes both long- and short-term trends. A previous study (Benson et al., 2020) showed that a combined WaveNet and long short-term memory (LSTM) model performed well in sunspot predictions, and concluded that Solar Cycle 25 would be slightly weaker than Solar Cycle 24.

The goal of this study is to predict Solar Cycle 25 through the deep learning approach, and determine what factors affect the prediction accuracy and what the optimal number of historical solar cycles are used to reliably and accurately predict the upcoming solar cycle. There is a tradeoff involved in the deep learning prediction approach: training the model with fewer numbers of historical solar cycles will result in a less accurate prediction of the multidecadal long-term trends in solar cycles. If more historical solar cycles are used for training the model, there is not enough observed data to train the model rigorously since large amounts of historical sunspot data would be required to make each prediction. This study will also explore the effects of data normalization methods on the final prediction accuracy.

One deep learning model that is ideal for this research question, involving time series predictions, is the LSTM model (Hochreiter and Schmidhuber, 1997). LSTM is a type of recurrent neural network (RNN), good at analyzing data over time, and can predict nonlinear patterns because it is capable of modeling both long-term and short-term dependencies in practice. It contains a cell state that transfers information from module to module throughout the entirety of the model. Information passed through the cell state is regulated by gates, which decides whether to add or remove information from the cell state. The sunspot number has multidecadal long-term trends in addition to the quasi-periodic 11-year cycles (Hathaway, 2015). One model that is useful in predicting long term trends is the convolutional neural network (CNN) (LeCun and Bengio, 1995). CNNs have performed better than RNNs in tasks involving modeling a time varying phenomenon and require less training time (Benson et al., 2020). The deep learning model used in this study is a combination of a one-dimensional convolutional neural network (1D CNN) and the LSTM model.

Dataset

The main dataset used in this study is the Kaggle dataset of observed monthly mean sunspot numbers from January 1749 to January 2021. The observed monthly mean sunspot data is updated to June 2022 using data from the Sunspot Index and Long-term Solar Observations (SILSO) website. The monthly mean sunspot number is a better dataset to use since there is less noise in this data than the daily sunspot dataset. This monthly mean dataset also extends further back than the daily sunspot dataset (1749 compared to 1818) and contains 3282 samples of monthly mean sunspot numbers. The average monthly mean sunspots in this dataset is ~82 and the corresponding standard deviation of the monthly mean sunspots is ~68. In this dataset, large noises are still seen near the peak of each solar cycle, hence the monthly mean data is further smoothed by taking the 13-month running average to reduce noises (Figure 1, Figure 2). Reducing noises is imperative to avoid overfitting with the deep learning prediction model and mistaking noise for important information about the solar cycle.

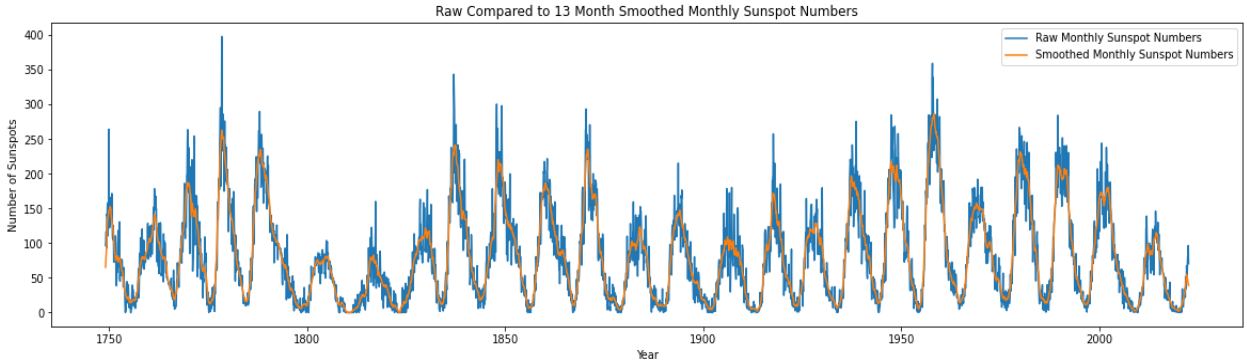

Figure 1. Time series of observed monthly mean (blue line) and 13-month smoothed (orange line) sunspot numbers.

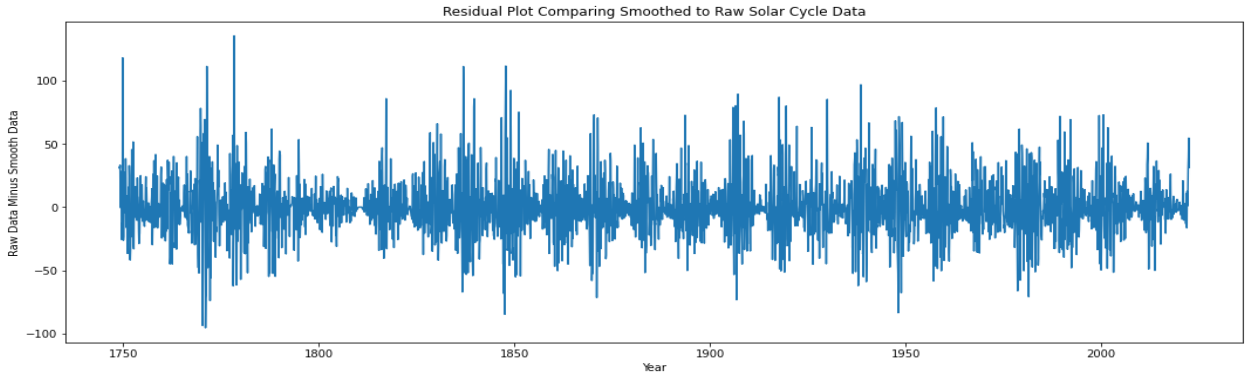

Figure 2. Residual sunspot number, i.e. the difference between raw monthly average sunspot numbers and 13-month smoothed sunspot numbers.

After the monthly mean sunspot data is smoothed, two types of data normalization, i.e. the MinMax normalization (Equation 1, Figure 3) and the Standard Deviation normalization (Equation 2, Figure 4), are evaluated to determine which type of data normalization is better for the deep learning model predictions of solar cycles. As will be discussed later, the MinMax normalization method (Equation 1) substantially reduces the error of the deep learning model's solar cycle predictions compared to the Standard Deviation normalization method (Equation 2). Hence most of this study is focused on using the MinMax normalization method (Equation 1).

Equation 1: MinMax normalization:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

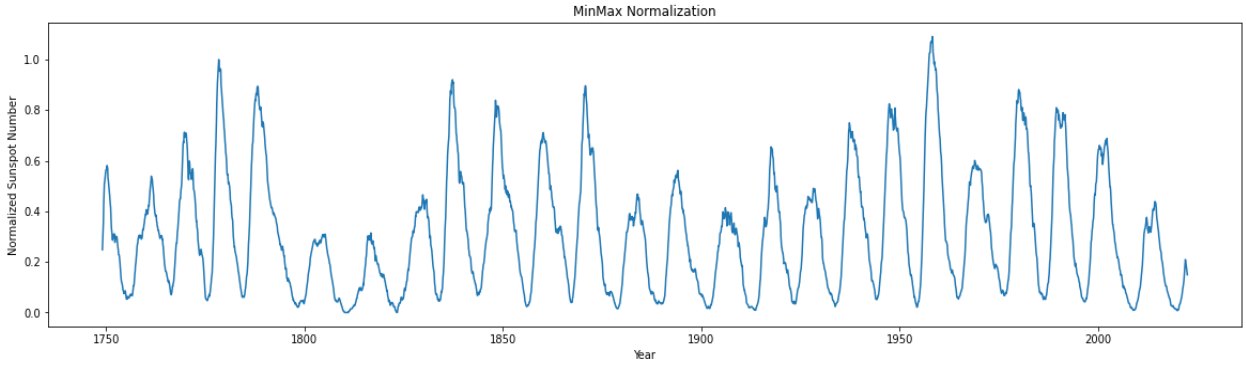

Figure 3. Sunspot dataset normalized using equation 1 (MinMax normalization).

Equation 2: Standard Deviation normalization:

$$x_{norm} = \frac{x - x_{mean}}{x_{std}}$$

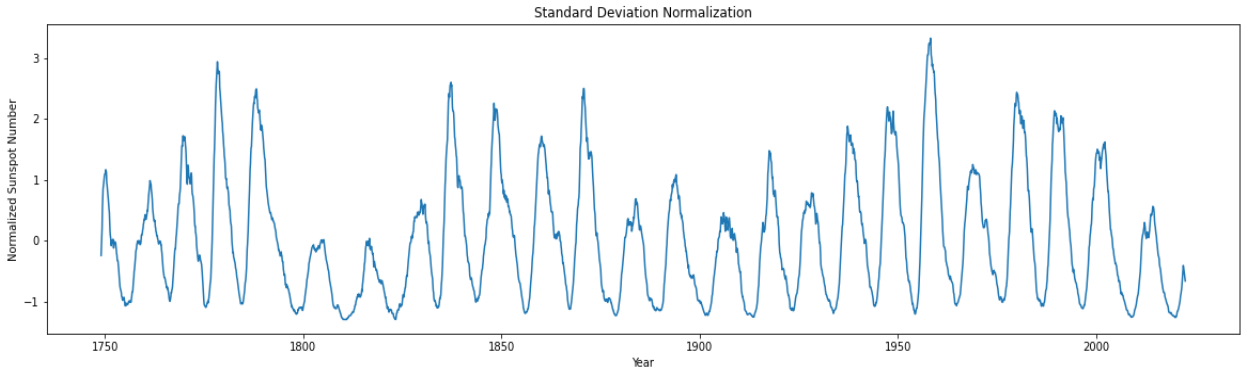

Figure 4. Sunspot dataset normalized using equation 2 (Standard Deviation normalization).

Model and Methodology

The deep learning model used in this study contains a 1-dimensional (1D) CNN layer connected to two LSTM layers, which is finally connected to a dense layer (Figure 5). The hybrid CNN-LSTM model combines the advantages of both CNN and LSTM for predicting time series cycles and trends. The CNN-LSTM model constructed for solar cycle predictions is supervised learning. The problem is a form of multi-step time series forecasting, i.e. using a consecutive number of historical solar cycles to predict one future solar cycle. Hence the target multi-step prediction output is monthly sunspot numbers for the duration of one solar cycle, or approximately the next eleven years (i.e. 132 months). The preprocessed 13-month smoothed and normalized sunspot data is first split into training data and validation data. For multi-step time series forecasts, the training and validation data cannot be split randomly. The split is chronological and designed to have equal sizes of segments for the training and validation data so that there is long enough validation data when using a large number of historical solar cycles to predict the future solar cycle. Each segment contains a pair of consecutive historical solar cycles and a target future solar cycle. The multi-step prediction code is adapted from the Kaggle "Tensorflow Time series LSTM Tutorial" written by Brooks for weather forecasts, using the corresponding functions from tensorflow.keras. The segments of historical solar cycles and the future solar cycle are created using the tensorflow library. The number of historical solar cycles (N) used to predict the future solar cycle determines the window size (number of months = $N \times 132$ months) of the historical part of the pair. This process was

repeated and each time the window is shifted forward by one month for the training/validation data. This is called the sliding window method, and it trains the model on each window of successive historical solar cycles to predict one future solar cycle.

The batch size of the training process is 20. The loss function used to train the model is mean absolute error (MAE) and the optimizer used is adam. The MAE of the model is evaluated to determine its prediction capability to match the target future solar cycle. MAE is not as sensitive to outliers as mean squared error (MSE), thus it is less affected by noises in the dataset. The kernel size for the 1D CNN is 10 for most of this study. When using the Standard Deviation normalization method for the sunspot data, the training loss explodes with the kernel size of 10 for the 1D CNN. Hence the kernel size of 20 is used for the 1D CNN when evaluating the two different data normalization methods. Early stopping is also implemented from tensorflow.keras.callbacks, which returns the model to a state with its lowest validation loss if a certain threshold number of epochs after that state all have a higher validation loss. In this case, that threshold number was 5 epochs, or a patience value of 5. This is intended to prevent overfitting and maximize the model's performance on validation data. The model is run in the Google Colaboratory environment.

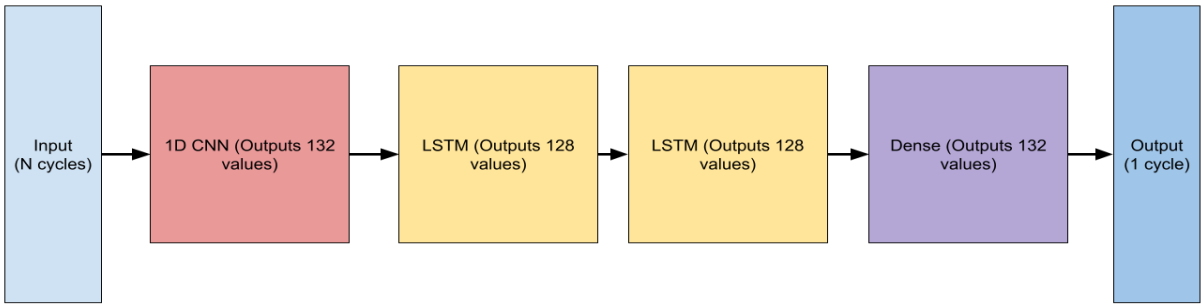

Figure 5. The diagram of the hybrid CNN-LSTM model structure.

Results and Discussion

Impact of Data Normalization

Figures 6 and 7 show that the MinMax normalization method (Equation 1) yields superior solar cycle prediction results compared to the Standard Deviation normalization method (Equation 2). With the latter method, the MAE for the validation data diverges rapidly and dramatically from the MAE for the training data (Figure 7), indicating that overfitting occurs with this normalization method. Both the training loss and validation loss are much larger for the Standard Deviation normalization than for the MinMax normalization (Figures 6 and 7). This is likely because the Standard Deviation normalization produces negative values (Figure 4), whereas the MinMax normalization keeps the data in the non-negative range between 0 and 1 (Figure 3). Since it is impossible to have negative monthly mean sunspot numbers, the Standard Deviation normalization (Equation 2) is not an appropriate way to normalize the sunspot data for this task. The Standard Deviation normalization method has been used in very recent studies for sunspot predictions through the deep learning approach (e.g. Prasad et al., 2022). The results shown here cast doubt on the accuracy of deep learning models' sunspot predictions using the Standard Deviation normalization method.

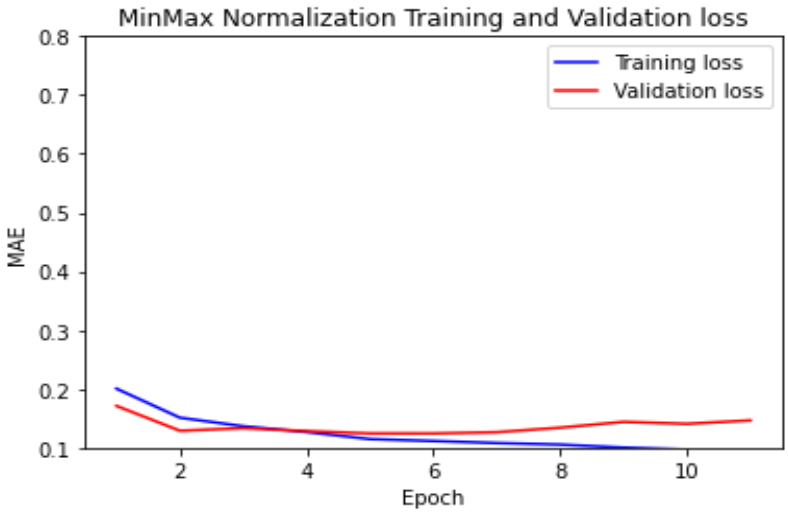

Figure 6. Training (blue) and validation (red) loss per epoch for the CNN-LSTM model trained on one historical solar cycle using the MinMax normalization.

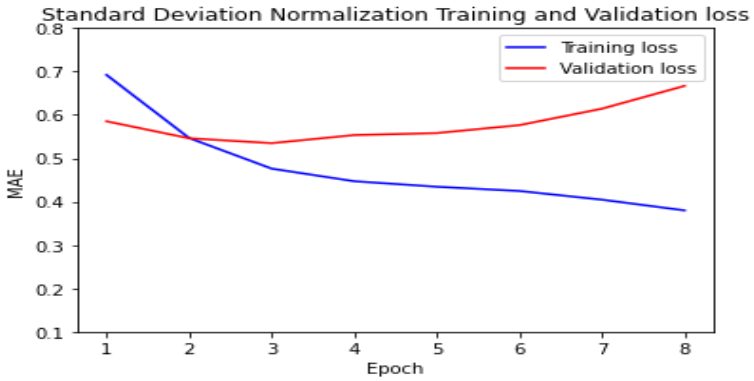

Figure 7. Training (blue) and validation (red) loss per epoch for the CNN-LSTM model trained on one historical solar cycle using the Standard Deviation normalization.

The 13-month smoothing is also effective in reducing the model's chance of overfitting the noises within the solar cycle dataset. Because LSTM is a model that depends on memory and patterns, it suppresses random noises in the output. As will be shown later in this section, the model's predictions match well with the 13-month smoothed true data. Thus, it is better to evaluate and interpret the model's predictions in terms of the 13-month smoothed data instead of the unsmoothed data so that the model's performance is not misinterpreted and underestimated because of the noises.

Impact of the Number of Historical Solar Cycles Used for Predicting the Future Solar Cycle

Figure 8 shows the grid search plot, i.e. the MAE of validation data as a function of the number of historical solar cycles (from 1 to 8) used for the CNN-LSTM model predictions. The best number of historical solar cycles used to predict a future solar cycle is 4 solar cycles (Figure 8), since it has the lowest overall MAE out of all the numbers (from 1 to 8) of the historical solar cycles considered. The metric displayed in Figure 8 is calculated using the same number of validation segments that is available for all models trained on different numbers (from 1 to 8) of historical solar cycles for a fair evaluation. Models trained on more (fewer) numbers of historical solar cycles have smaller (larger) sizes of available training and validation segments. For example, the model trained on 8 historical solar cycles only has 519 available training and validation segments. For this reason,

models trained on the number larger than 8 historical solar cycles are not considered, as there are not large enough segments for training and validation. Using different available validation segment sizes for the model evaluation will skew the accuracy metrics towards models trained on the lower numbers of historical solar cycles. By using the same size of validation segments (i.e. the last 519 segments of the validation data) for all models as that is available for the model trained on the largest number (8) historical solar cycles, the bias towards the models trained on the lower numbers of historical solar cycles in the accuracy metrics is reduced.

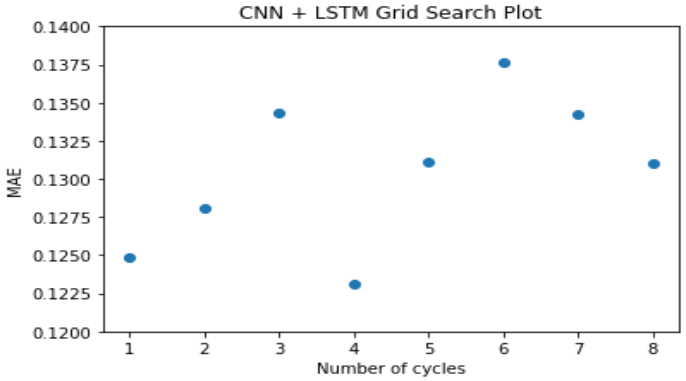

Figure 8. Grid Search Plot, i.e. the MAE of validation data as a function of the number of historical solar cycles (from 1 to 8) used for the CNN-LSTM model predictions.

It is difficult for a model with less than 4 historical cycles to reproduce the longer multidecadal solar cycle trends in their predictions (Benson et al., 2020). So, even if the model trained on 1 historical solar cycle have similar MAE to the model trained on 4 historical cycles (Figure 8), it will struggle to replicate the multidecadal long term trends in solar cycles (Figure 1) and will not perform as well as the model trained on 4 historical solar cycles when predicting the multidecadal trends in the peak amplitudes of solar cycles. Among models trained on at least 4 historical cycles (from 4 to 8), the model trained on 4 historical cycles clearly performs the best (Figure 8) and is used to predict Solar Cycle 25. The model trained on 4 historical solar cycles effectively predicts longer multidecadal trends present in solar cycles while having enough history-future segments to train more rigorously than models that used larger numbers of historical solar cycles. As seen in Figure 9, validation loss eventually begins to gradually increase as the training loss continues to decrease. This proves the necessity of early stopping in preventing overfitting later in the model's training process. As depicted in Figures 10 and 11, the model trained on 4 historical solar cycles can predict well how powerful the peaks of solar cycles 23 and 24 are and when they occur.

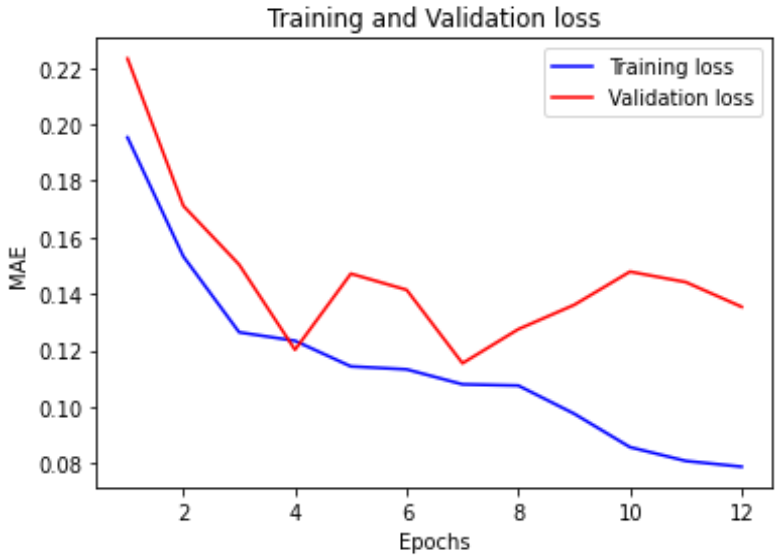

Figure 9. Training loss (blue line) compared to validation loss (red line) per epoch for the CNN-LSTM model trained on four historical solar cycles.

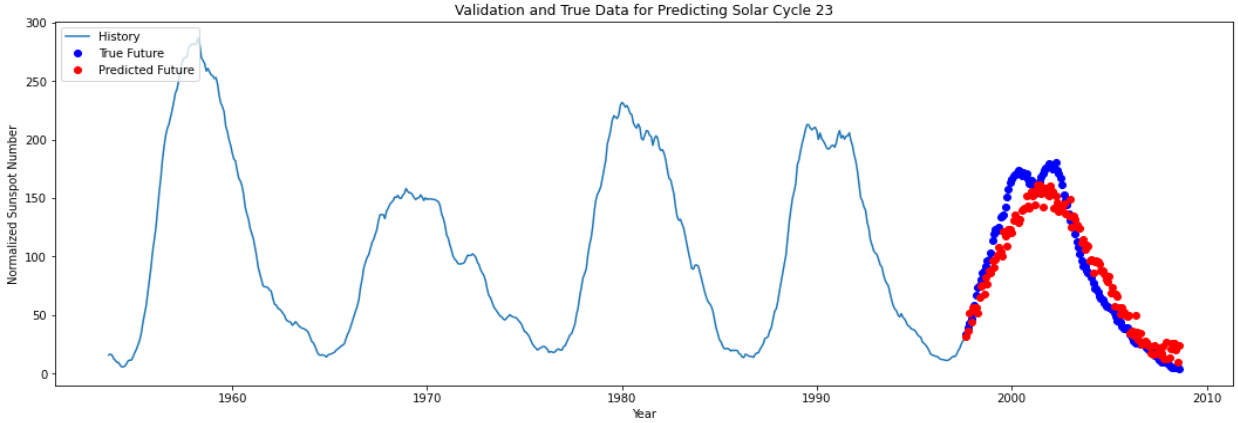

Figure 10. The 4-cycle based CNN-LSTM predictions (red dot) compared to the true data (blue dot) for solar cycle 23.

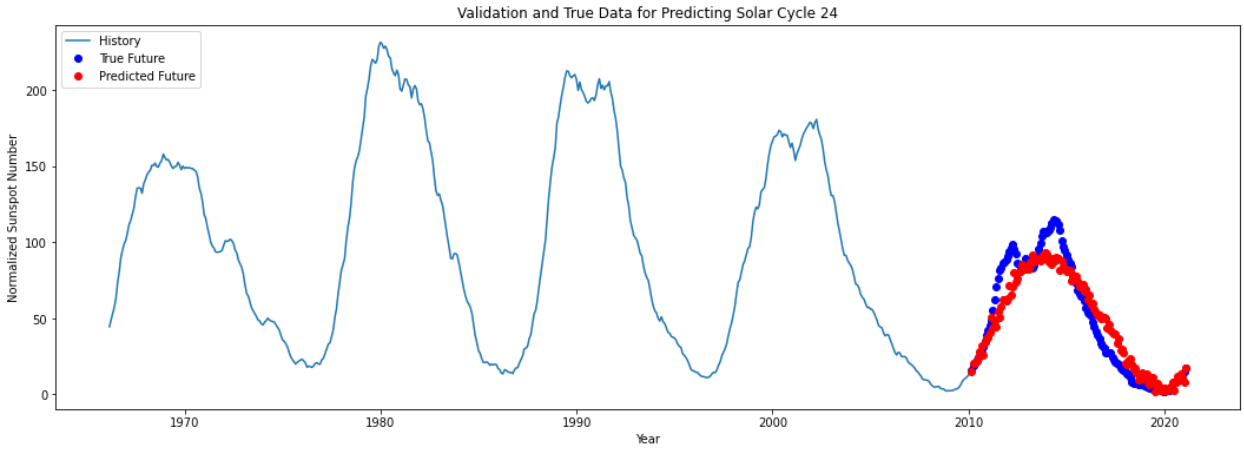

Figure 11. The 4-cycle based CNN-LSTM predictions (red dot) compared to the true data (blue dot) for solar cycle 24.

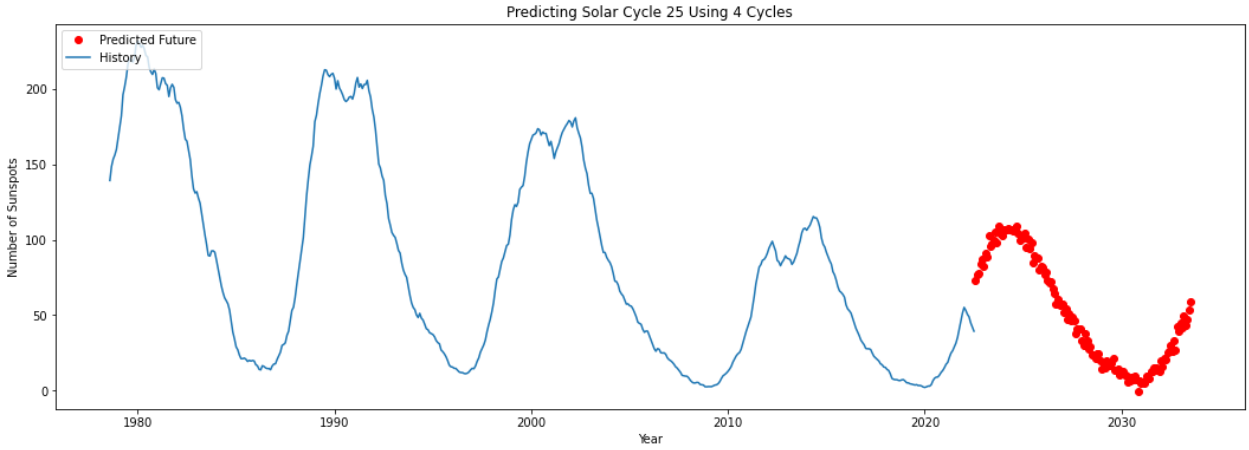

Figure 12. The predictions for solar cycle 25 (red dot) using 4 historical solar cycles (blue line).

Predictions for Solar Cycle 25

The CNN-LSTM model trained on 4 historical solar cycles predicts the peak of Solar Cycle 25 at around 110 ± 22 sunspots, which will occur at approximately from August 2023 to July 2024 (Figure 12). Here, the uncertainty of 22 sunspots is one standard deviation of the residual sunspot number (Figure 2). The predicted peak is rounded instead of being sharp, which explains why the peak spans a relatively long period of time between approximately August 2023 and July 2024. This result shows that Solar Cycle 24 (with a 13-month smoothed peak of around 115 sunspots) and Solar Cycle 25 are similar in the peak strength. This result has a similar predicted peak amplitude as that released by the NOAA/NASA co-chaired international panel, but here the predicted peak occurs earlier than that consensus prediction. A study using solar magnetic activity cycle data (McIntosh et al., 2020) predicted that Solar Cycle 25 could be one of the strongest solar cycles to date, and markedly more powerful than Solar Cycle 24. Another very recent paper (Prasad et al., 2022) using the LSTM model trained on 10 historical solar cycles with the Standard Deviation data normalization method also predicted a very powerful Solar Cycle 25 peak. While some might look at how current sunspot observations are outpacing Solar Cycle 25 predictions released by the NOAA/NASA co-chaired international panel and speculate the peak of Solar Cycle 25 to be much higher than that of Solar Cycle 24, this is not necessarily the only scenario. Alternatively, it could be that the sunspot number of Solar Cycle 25 simply rises earlier than the

consensus prediction. Recent observed sunspot numbers may indicate the possibility of an early, but not necessarily powerful peak in Solar Cycle 25.

This study yields similar Solar Cycle 25 peak predictions to another study employing a WaveNet and LSTM model to forecast 1 future solar cycle using 4 historical solar cycles (Benson et al., 2020), which predicted a Solar Cycle 25 peak of ~106 sunspots. Benson et al. 2020 used the unsmoothed monthly mean sunspot data, while this study used the 13-month smoothed data. The two studies also employ different metrics to evaluate the performance of the models: Benson et al. 2020 used the root mean squared error (RMSE) while this study used MAE. This study suggests that it is better to interpret the LSTM model's predictions in terms of the 13-month smoothed data instead of unsmoothed monthly mean data, because LSTM is a model designed to depend on memory and suppress random noises in the output. Indeed, the predicted solar cycles in Benson et al., 2020 are much more smoothed and less noisy than the unsmoothed monthly mean sunspot data. Additionally, this study investigated the impacts of data normalization methods and the number of historical solar cycles used to train the models on the prediction accuracy. Benson et al. 2020 suggested that at least 4 historical solar cycles should be used to train the deep learning prediction model. This study suggested that, among models trained on at least 4 historical cycles (from 4 to 8), the model trained on 4 historical cycles clearly performs the best.

Conclusion

In this study, the impacts of factors such as the type of data normalization and the number of historical solar cycles used in predicting the future solar cycle are investigated. A hybrid CNN-LSTM model was constructed, which involved a 1D CNN layer, two LSTM layers, and one dense layer. It is clearly shown that the MinMax normalization produces better prediction results than the Standard Deviation normalization. The 13-month smoothed sunspot dataset is more compatible with the reduction of noises in the predictions of the CNN-LSTM model. Using the Standard Deviation normalization, the model could predict negative sunspot values, which would never happen in the real-world solar cycles, and the loss is significantly higher than using the MinMax normalization. This study also suggests that the best number of the historical solar cycles used in predicting the future solar cycle is 4 solar cycles. Using 4 historical cycles can cover the multidecadal long-term trends in solar cycles that are not covered in models trained with much smaller historical cycle numbers. Here the CNN-LSTM model trained on 4 historical solar cycles predicts that Solar Cycle 25 has a 13-month smoothed peak amplitude similar to that of Solar Cycle 24, occurring approximately between August 2023 and July 2024.

As the solar cycle is a periodic time series with long-term trends, the methods utilized in this paper can be applied to predict other time series with similar features. Future studies could investigate multiple variables by incorporating data such as x-ray and radio wave emissions from the Sun. Emissions like these are also related to the solar cycles and have a similar periodicity, and they will paint a more complete picture of solar cycles.

Acknowledgments

I thank Kaggle and the Sunspot Index and Long-term Solar Observations (SILSO) website for providing the monthly average sunspot dataset (https://www.kaggle.com/datasets/robervalt/sunspots, https://www.sidc.be/silso/datafiles) used in this paper. I also thank my mentor (name listed in the metadata) for mentoring me on this project and the Inspirit AI 1:1 mentorship program (https://www.inspiritai.com/returning-students) for providing me this research opportunity.

References

Benson, B., Pan, W. D., Prasad, A., Gary, G. A., & Hu, Q. (2020). Forecasting solar cycle 25 using deep neural networks. Solar Physics, 295(5), 1-15. https://doi.org/10.1007/s11207-020-01634-y

Brooks, N. Tensorflow Time series LSTM Tutorial. https://www.kaggle.com/code/nicapotato/keras-timeseries-multi-step-multi-output, retrieved July 2022.

Hathaway, D. H. (2015). The solar cycle. Living reviews in solar physics, 12(1), 1-87. https://doi.org/10.1007/lrsp-2015-4

Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural Comput. 9(8), 1735. https://doi.org/10.1162/neco.1997.9.8.1735

Jyothi, S. A. (2021). Solar superstorms: planning for an internet apocalypse. In Proceedings of the 2021 ACM SIGCOMM 2021 Conference (pp. 692-704). https://doi.org/10.1145/3452296.3472916

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), 1995. http://www.iro.umontreal.ca/~lisa/pointeurs/handbook-convo.pdf

McIntosh, S. W., Chapman, S., Leamon, R. J., Egeland, R., & Watkins, N. W. (2020). Overlapping magnetic activity cycles and the sunspot number: forecasting sunspot cycle 25 amplitude. Solar Physics, 295(12), 1-14. https://doi.org/10.1007/s11207-020-01723-y

Moore, T. E. (2020). The cosmic timeline of heliophysics: A declaration of significance. Perspectives of Earth and Space Scientists, 1(1), https://doi.org/10.1029/2020CN000137

National Research Council. (2013). Solar and Space Physics: A Science for a Technological Society. Washington, DC: The National Academies Press. https://doi.org/10.17226/13060

Prasad, A., Roy, S., Sarkar, A., Panja, S. C., & Patra, S. N. (2022). Prediction of solar cycle 25 using deep learning based long short-term memory forecasting technique. Advances in Space Research, 69(1), 798-813. https://doi.org/10.1016/j.asr.2021.10.047