SOLAR SPECTRUM - UNDERSTANDING SOLAR ACTIVITY PATTERNS WITHIN A SOLAR CYCLE USING MACHINE LEARNING TECHNIQUES

Project Reference No.: *47S_BE_3481*

College ***:*** *Bangalore Institute of Technology, Bengaluru*
Branch ***:*** *Department of Computer Science & Engineering*
Guide(s) ***:*** *Dr. M S Bhargavi*
Student(S) ***:*** *Ms. Deeksha Deepak*
Ms. Devika U H
Ms. Disha K Sanketh
Ms. Gayathri R

Keywords:
Solar cycle, Sunspots, and Coronal Mass Ejections, Solar flares, Machine Learning, LSTM, GRU.

Abstract:

Solar activity is a critical factor shaping space weather, influencing satellite communications, power systems, and even Earth's climate. Recognizing the need for accurate predictions in this domain, the Solar Activity Prediction System is a data science engineering project that aims to harness advanced machine learning techniques. The primary focus is on developing a robust prediction model utilizing the Gated Recurrent Unit (GRU) architecture, chosen for its advantages over the more traditional Long Short-Term Memory (LSTM) model.
The overarching objective of this project is to enhance the accuracy of solar activity predictions by shifting the emphasis from conventional solar activity indicators to predicting the actual solar activity itself. This necessitates a thorough understanding of historical data encompassing various parameters associated with solar activity, such as sunspots and solar flares. The chosen approach involves training and validating time-based models, specifically the GRU and LSTM models. The overarching objective of this project is to enhance the accuracy of solar activity predictions by shifting the emphasis from conventional solar activity indicators to predicting the actual solar activity itself. This necessitates a thorough understanding of historical data encompassing various parameters associated with solar activity, such as sunspots and solar flares. The project also emphasizes graphical representation to visually convey the disparities between historical data and model predictions, offering a comprehensive analysis of the solar activity forecasting system's performance. The chosen approach involves training and validating time-based models, specifically the GRU and LSTM models. This project seeks to contribute to advancements in space weather prediction and the resilience of technological infrastructures against the impacts of solar activity.

Introduction:

The Sun has a massive impact on our planet and space behavior. Many changes, disturbances, and phenomena keep occurring in and around the Sun's atmosphere, generally referred to as Solar activity. The prediction and classification of Solar activity and Solar phenomena in general with methods of machine learning (ML) has become an active field of research in astrophysics.

Solar activity, characterized by phenomena such as solar flares and sunspots, exhibits complex and dynamic patterns. Accurate prediction of these patterns is crucial for enhancing the efficiency and reliability of solar energy systems. Most of the effort within the literature has been devoted to the prediction of solar flares, rather than to the classification of different types of solar activity. These efforts are most often based on photo-spheric vector-magnetic field data from the Solar Dynamics Observatory.

Nevertheless, instead of defining solar activity from a magnetic perspective, we have chosen to categorize the activity by monitoring directly observed spectral responses. Such input data would be model free, and not rely on tenuous extrapolations of photo-spheric magnetic data into the corona, like the works on Spectro-polarimetric diagnostics and inversion methods using ML.

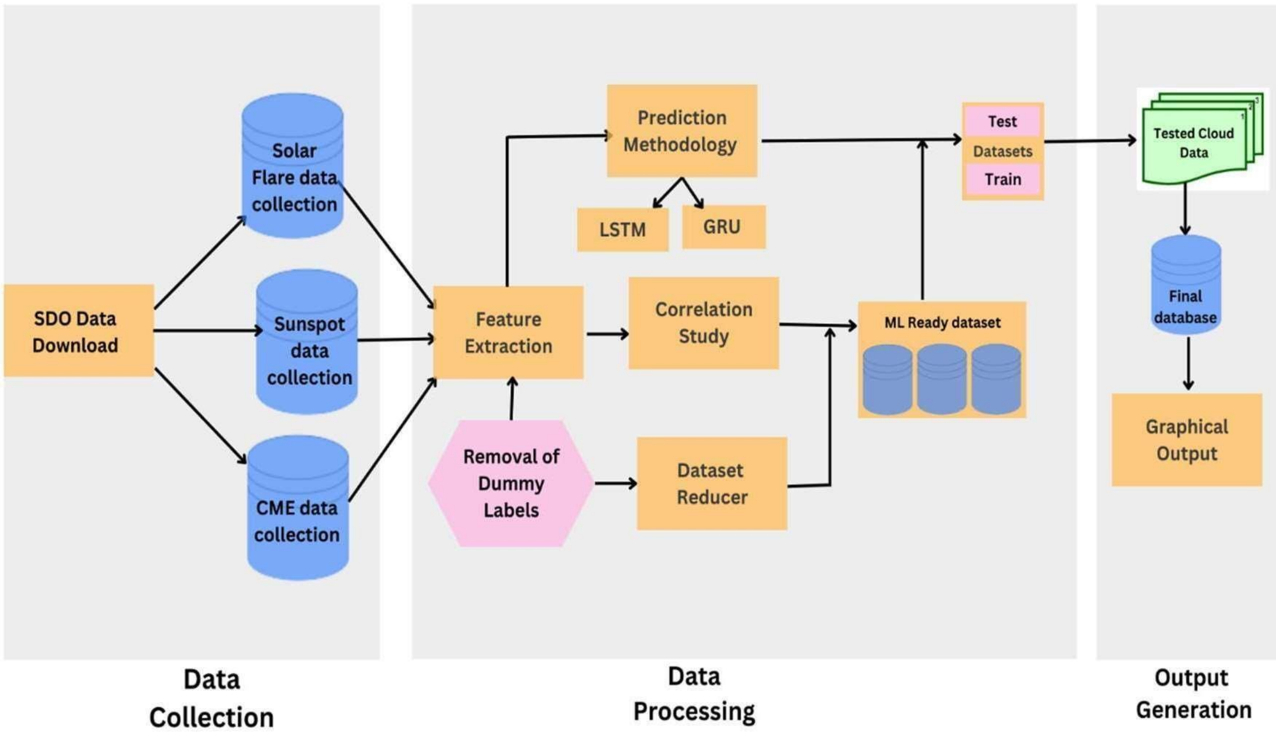
To develop a predictive model that can forecast the solar activities, the project aims to enhance prediction accuracy by shifting focus from solar activity indicators to actual solar activity, employing time-based models such as GRU and LSTM. A principle finding of this research is that the classification performance on compressed and uncompressed data is comparable under our architecture, implying the possibility of large compression rates for relatively low degrees of information loss. With an exponential growth of solar data, fast algorithms that can guide researchers to specific solar activity becomes a very useful tool.

Objectives:

The objective of this project is to develop a robust machine learning model capable of predicting solar activity with high accuracy. The primary objectives of this project are:

- To implement a predictive model capable of forecasting the solar activities.
- To serve as a forecasting mechanism that generates predictions for the occurrence and intensity of upcoming solar activities.
- To accomplish this, train the selected model using historical data and validate its performance using suitable metrics.
- The final solution should be able to identify patterns and indicators that precede solar activity and mitigates the potential impact of the same on earth's system.

Methodology:



The methodology used involves a three-stage process related to solar flare prediction or space weather forecasting using machine learning techniques. This intricate process includes a series of steps aimed at collecting, processing, and generating insights from data to enhance our understanding of solar activities.

The first stage of the process focuses on Data Collection. It involves gathering crucial information from various sources such as the Solar Dynamics Observatory (SDO), which provides high- resolution sun data. Additionally, the process includes collecting data on sunspots, which are temporary dark spots on the Sun's surface indicating solar activity, as well as data on solar flares and Coronal Mass Ejections (CMEs). These components serve as the foundation for further analysis and prediction of solar phenomena.

Moving on to the Data Processing stage, the flowchart outlines several key tasks. Firstly, it involves Feature Extraction, where relevant features are identified and extracted from the collected data to facilitate analysis. Subsequently, the process includes the Removal of Dummy Labels to eliminate irrelevant labels and focus on meaningful data points. Moreover, a Correlation Study is conducted to understand the relationships between different features in the dataset. The flowchart also highlights the importance of Dataset Reduction for efficiency and performance optimization, followed by preparing the dataset to be machine learning-ready for subsequent analysis.

Thus time-series forecasting models like recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, or Gated Recurrent Units (GRUs) are used for prediction and for binary classification tasks (flare or no flare), models like Random Forests, Support Vector

Machines (SVM), or neural networks can be effective. We ascribe data taken T hours prior to a flare to the positive class and all other data to the negative class in the prediction model. Ultimately, we are interested developing a real-time predictive model that predicts future solar activity given current solar data. We simulate this process by splitting our data into two subsets. The first subset, or training set, contains the data our algorithm will learn from. The other subset, or testing set, is used to evaluate our algorithm. In the machine learning literature, this is known as cross-validation. Here we always use a test set size that constitutes approximately 20% of the data.

The final stage of the process is Output Generation, which involves several critical steps. It includes splitting the dataset into Test Datasets for training and testing machine learning models. The results obtained from testing are stored in a Final Database for further analysis and reference. Additionally, the flowchart emphasizes the importance of presenting the results in a Graphical Output format to facilitate better understanding and interpretation of the findings.

Results and Conclusion:

In the analysis of solar activity, both the GRU and LSTM models demonstrated minimal overfitting. However, an interesting distinction emerged between the two models: while the LSTM model produced predicted and validated values that closely matched each other, the GRU model generated a horizontal line, indicative of the mean sunspot number's average. This difference suggests that the LSTM model may have better captured the temporal dynamics of sunspot activity compared to the GRU model. Multivariate solar flare prediction showcased varied accuracies, with LSTM at 72.76% and GRU at 71.43%, while combined models reached up to 83.89%. Additionally, binary classification of CMEs using LSTM and GRU models resulted in an AUC of 55% and 47%, respectively. These findings offer significant perspectives into the efficacy of RNN architectures in solar activity forecasting, useful for advancing solar activity forecasting systems.

By leveraging advanced algorithms and data-driven approaches, researchers have the potential to enhance our comprehension of solar dynamics, contributing to improved predictions of space weather and its impacts on Earth. The insights gained from this research can not only enhance our fundamental understanding of the sun but also have practical implications for space exploration, satellite communication, and power grid management.

Furthermore, as solar cycles play a crucial role in shaping the sun's behavior, the integration of machine learning into solar studies can pave the way for more accurate forecasting models. This has the potential to mitigate the effects of solar events on technology-dependent systems on Earth, ensuring better preparedness and resilience.

Innovations in the Project:

The project aims to compare the effectiveness of two prominent recurrent neural network (RNN) architectures, Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM), in forecasting solar activities. The study focuses on three primary solar phenomena: sunspots, solar flares, and coronal mass ejections (CMEs). For sunspots prediction, a univariate dataset is employed, featuring monthly mean sunspot counts as the target variable and month as the independent variable. In the realm of solar flares prediction, a multivariate dataset is utilized, incorporating various features related to solar activity. The study examines LSTM, GRU, and combinations thereof, including models with and without recurrent dropouts. Additionally, the study addresses the classification of coronal mass ejections (CMEs), often associated with solar flares. Both LSTM and GRU models are employed for binary classification based on independent variables.

Scope for Future Work:

The project scope encompasses a meticulous approach to Solar Activity Prediction through Machine Learning Algorithms, commencing with the collection and preprocessing of historical solar activity data and relevant atmospheric conditions.
Ultimately, we are interested developing a real-time predictive model. In other words, we are interested in predicting future flaring activity given current solar data. A principle finding of this research is that the classification performance on compressed and uncompressed data is comparable under our architecture, implying the possibility of large compression rates for relatively low degrees of information loss.

- Future work includes using the superior model, whether it be LSTM or GRU, to make predictions regarding the current solar cycle.
- Investigating whether solar flare occurrences were accompanied by Coronal Mass Ejections(CMEs) or not is also part of the future work.
- Improving accuracies through more detailed Exploratory Data Analysis (EDA) is anotheraspect to be considered for enhancement.