International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 10-No.2 (2024) pp. 189-199
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Observation of the Long-Term Relationship Between Cosmic Rays and Solar Activity Parameters and Analysis of Cosmic Ray Data with Machine Learning

Ahmet POLATOĞLU[1]*

[1] Atatürk University, Faculty of Sciences, Department of Astronomy and Space Sciences, 25240, Erzurum-Türkiye
* **Corresponding Author Email:** ahmet.polatoglu@atauni.edu.tr - **ORCID:** 0000-0002-6562-8566

Article Info:

DOI: 10.22399/ijcesen.324
Received : 21 May 2024
Accepted : 06 June 2024

Keywords

Cosmic Rays (CR)
Machine Learning (ML)
Regression
Solar Activity
Space Weather

Abstract:

Understanding the complex interplay between solar activity and cosmic ray intensity is crucial for unraveling the mysteries of space weather and its impacts on Earth's environment. In this study, I investigate the relationships between solar activity parameters and cosmic ray intensity using a comprehensive dataset obtained from the LASP Interactive Solar IRradiance Datacenter (LISIRD) and the OULU neutron database. Through data visualization, correlation analysis, and machine learning techniques, I analyze decades of solar and cosmic ray data to discern patterns, trends, and correlations over time. Findings reveal significant correlations between solar activity parameters such as the sunspot number (SSN), Mg II Index, and various radio flux (RF) measurements at different wavelengths, with cosmic ray intensity. Notably, I observe a strong inverse correlation between SSN and RF at 30 cm with a value of -0.82, indicating the influence of solar activity on modulating cosmic ray flux reaching Earth. Machine learning models, including Gradient Boosting Machines (GBM) and Artificial Neural Networks (ANN), are employed to predict cosmic ray intensity, achieving promising results. Furthermore, regularization techniques such as Ridge and Lasso regression are utilized to mitigate overfitting and improve prediction performance. My study underscores the importance of integrating diverse datasets and employing advanced analytical approaches to enhance our understanding of solar-cosmic interactions and their implications for space weather forecasting. These insights have implications for various fields, from astrophysics to atmospheric science, and contribute to ongoing efforts aimed at deciphering the complexities of cosmic phenomena and their impacts on Earth's environment.

1. Introduction

The Sun, our nearest star, is a dynamic and complex system that profoundly influences the heliosphere and planetary environments within it [1]. Understanding solar activity and its interactions with cosmic rays (CR) is crucial for a range of scientific and practical applications, from climate modeling to space weather forecasting. In recent years, machine learning (ML) techniques have emerged as powerful tools for uncovering patterns and relationships in complex datasets, offering new avenues for exploring the intricate dynamics between solar activity and CRs over long periods [2-4].

Solar activity encompasses various phenomena, including sunspots, solar flares, and coronal mass ejections (CMEs). These activities are often quantified by parameters such as the sunspot number, solar radio flux (F10.7 index), and geomagnetic indices (e.g., Kp index). Solar activity follows an approximately 11-year cycle, characterized by alternating periods of high and low activity. During solar maxima, increased solar activity leads to more frequent and intense solar events, whereas solar minima are marked by a relative calm [5]. CRs are high-energy particles originating from outer space that travel at nearly the speed of light. They are a form of ionizing radiation and consist primarily of protons, atomic nuclei, and a small fraction of heavier elements and electrons [6]. CRs are modulated by solar activity as they travel through the heliosphere. The solar wind, a stream of charged particles emanating from the Sun, creates a bubble-like region that influences the

propagation of CRs. During periods of high solar activity, the enhanced solar wind and magnetic field provide a more effective shield against CR, leading to a decrease in their intensity observed at Earth. Conversely, during solar minima, reduced solar wind pressure allows more CRs to penetrate the heliosphere [7-8].

The relationship between solar activity and CR is well-documented but remains complex and multifaceted. Traditional methods of analysis have provided significant insights, yet they often struggle to account for the non-linear and multivariate nature of the data. Here, machine learning offers a promising approach to overcome these limitations. ML algorithms can handle large datasets with numerous variables, identify intricate patterns, and make predictions based on historical data. This makes them particularly well-suited for analyzing the long-term relationship between solar activity parameters and CRs. Machine learning encompasses a variety of techniques, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning algorithms, such as decision trees, support vector machines (SVMs), and neural networks, are trained on labeled datasets and are used to predict outcomes or classify data based on input features [9]. In the context of solar activity and CRs, supervised learning can help predict CR intensity based on observed solar activity parameters. Unsupervised learning algorithms, such as clustering and dimensionality reduction techniques, can uncover hidden structures in the data without predefined labels. These methods can be used to identify natural groupings or trends in solar and CR data. The integration of machine learning into the study of solar activity and CRs involves several key steps. First, it requires the compilation of extensive datasets covering various solar activity parameters and CR measurements over long periods. These datasets often come from multiple sources, including ground-based observatories, satellite missions, and space probes. Data preprocessing is crucial to ensure quality and consistency, involving tasks such as normalization, outlier detection, and handling missing values [10]. Next, feature selection and engineering play a critical role in improving model performance. Model selection and training follow, where different ML algorithms are evaluated to determine the best performing ones for the task. This involves training models on historical data, tuning hyperparameters, and validating their performance using techniques such as cross-validation. The chosen models can then be tested on unseen data to assess their predictive accuracy and generalization capability. Interpretability and deployment of the models are crucial for practical applications. While ML models can provide accurate predictions, understanding the rationale behind their decisions is essential for gaining scientific insights and ensuring their reliability in real-world scenarios. The application of machine learning to study the long-term relationship between solar activity and CRs holds significant promise. It can lead to improved predictive models that enhance our ability to forecast space weather events, which is vital for protecting satellites, astronauts, and ground-based technological infrastructure from cosmic ray-induced disruptions. Furthermore, it can deepen our understanding of the solar-terrestrial connection, contributing to broader scientific knowledge about the Sun's influence on our solar system. The intersection of solar physics, CR research, and machine learning represents a dynamic and evolving field with substantial potential. By leveraging the strengths of ML techniques, researchers can uncover new insights into the complex interactions between solar activity and CRs, paving the way for advancements in both fundamental science and practical applications [11]. As data availability and computational power continue to grow, the integration of machine learning into this domain will likely become increasingly sophisticated, driving further discoveries and innovations. In this study, cosmic ray and Solar activity parameters (SSN, Mg II Index, RF 3.2 cm, RF 8 cm, RF 10.7 cm, RF 15 cm, RF 30 cm and Lyman-alpha) between 1979-2024 were used. Firstly, time series graphs were drawn with these data and then the relationship between them was examined with correlation analysis. Long-term relationships between solar energy release and cosmic rays were examined using advanced machine learning techniques. Specifically, Gradient Boosting Machines (GBM) and Artificial Neural Networks (ANN) were used to model this relationship and GridSearchCV for optimal hyperparameter tuning. In addition, Linear Regression and regularization techniques such as Ridge Regression and Lasso Regression have been applied to compare and validate the findings. These methodologies allow for a comprehensive analysis, leveraging both the predictive power of complex models and the interpretability of linear approaches

2. Data and Methods

2.1. Data

The LASP Interactive Solar IRradiance Datacenter (LISIRD) is an online platform designed to facilitate access to solar data for researchers in the

field of heliophysics. Its primary objective is to streamline the process of solar data discovery, visualization, and retrieval by offering a comprehensive collection of datasets sourced from various space missions, instruments, models, and research laboratories. LISIRD is committed to enhancing the accessibility and usability of solar data through an intuitive user interface, comprehensive metadata, interactive plotting functionalities, and an extensive database of available datasets. Solar parameters taken from LISRID are such as the Sunspot Number (SSN), Mg II Index, and various radio flux (RF) measurements at different wavelengths (3.2 cm, 8 cm, 10.7 cm, 15 cm, and 30 cm), along with Lyman-alpha emissions, are crucial indicators of solar activity and its impact on space weather. The SSN reflects the number of visible sunspots and is a primary measure of solar activity cycles. The Mg II Index, derived from the ultraviolet emissions of the magnesium ion, provides insights into the solar chromosphere's conditions. Radio flux measurements at different wavelengths, particularly the 10.7 cm flux, serve as proxies for solar activity levels, correlating with sunspot numbers and magnetic activity. Lyman-alpha emissions, a specific ultraviolet light emitted by hydrogen atoms, indicate the amount of solar ultraviolet radiation reaching the Earth, influencing the ionization levels in the Earth's upper atmosphere. Collectively, these parameters help scientists understand and predict solar phenomena and their effects on the Earth's space environment [12]. CR data were taken from the OULU neutron database. The OULU neutron database serves as a vital repository for cosmic ray data, offering researchers a wealth of information to explore the intricacies of these high-energy particles. With meticulous collection methods and comprehensive storage, this database provides a robust foundation for studies spanning various fields, from astrophysics to atmospheric science. By accessing this repository, scientists can delve into the mysteries of cosmic rays, unraveling their origins, behaviors, and impacts on our planet and beyond. This data from the OULU neutron database stands as a cornerstone for advancing our understanding of these elusive cosmic phenomena, guiding future research endeavors and illuminating the secrets of the cosmos [13].

2.2. Data Visualization and Correlation Analysis

The datasets span decades, providing a comprehensive temporal range necessary for long-term analysis. Following the transfer of the data, the initial processing steps will be to handle missing values, normalize the data, and align the time series to ensure that all measures are maintained. The data sets are then converted into a unified time series format to combine subsequent analyses.

Time series graphs for each solar activity parameter and cosmic ray data were plotted to visualize trends, periodicities, and anomalies over time. This visual examination is crucial for understanding the underlying patterns and guiding further statistical and machine learning analyses.

To explore the relationships between solar activity parameters and cosmic ray intensity, we performed a correlation analysis using the Pearson correlation coefficient, defined as equation 1 [14].

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x^2)][n(\sum y^2) - (\sum y^2)]}} \quad (1)$$

where n is the number of observations, and x and y are the variables being compared. The resulting correlation matrix highlighted significant correlations, which informed the feature selection for subsequent modeling.

The performance of each model was evaluated using metrics such as Mean Squared Error (MSE) and R^2. MSE is represented by $\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2$ and R^2 is represented by $1 - MSE/\sum_{i=1}^{n}(y-\bar{y})^2$. Here, y is the actual value and $\hat{y}$ is predicted value of the *i-th* observation. $\bar{y}$ is the mean of the actual values. Cross-validation was employed to ensure the robustness and generalizability of the models. Comparisons were made to identify the most effective model for predicting cosmic ray intensity based on solar activity parameters [15].

2.3. Machine Learning Models

Gradient Boosting Machines (GBM) stand out as a robust machine learning approach employed for both regression and classification purposes. GBM constructs a series of decision trees step by step, with each subsequent tree aiming to rectify the mistakes of its predecessors. By amalgamating the forecasts of numerous weaker learners (usually shallow trees), GBM crafts a potent predictive model. The implementation of GBM is facilitated through the scikit-learn library, wherein decision trees are sequentially aggregated, each endeavoring to rectify the errors introduced by its precursors. The prediction $\hat{y}$ is given by equation 2.

$$\hat{y} = \sum_{m=1}^{M} \gamma_m h_m(x) \quad (2)$$

where $h_m(x)$ is the *m-th* weak learner, and γ_m is its corresponding weight. Hyperparameters such as learning rate, number of estimators, and maximum depth of trees were optimized using *GridSearchCV*, ensuring the model's robustness and generalization capability. GridSearchCV is a hyperparameter optimization technique provided by the scikit-learn library in Python. It is used to find the best combination of hyperparameters for a given machine learning model by systematically searching through a specified parameter grid [16,17].

Artificial Neural Network (ANN) are another important models for machine learning tasks. They consist of layers of neurons, each performing a weighted sum of its inputs followed by an activation function. The network learns by adjusting its weights and biases to minimize a loss function through backpropagation and gradient descent, enabling it to model complex, non-linear relationships in the data. An ANN is created using TensorFlow. The network architecture includes an input layer, multiple hidden layers containing neurons, and an output layer. Backpropagation and gradient descent were used to minimize the loss function and optimize the network parameters [18].
Linear regression was applied to model the linear relationship between solar activity parameters and cosmic ray intensity. The regression equation is given by equation 3 [19].

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (3)$$

where y is the dependent variable (cosmic ray intensity), x_i are the independent variables (solar activity parameters), β_i are the coefficients, and ϵ is the error term. The Ordinary Least Squares (OLS) method was used to estimate the coefficients.

2.4. Regularization Techniques

Ridge regression is a type of linear regression that's particularly useful when dealing with multicollinearity, which is when independent variables in a regression model are highly correlated. To address multicollinearity and overfitting, Ridge Regression was employed. The ridge regression coefficients β_{ridge} are given by equation 4. In the equation, λ is the regularization parameter controlling the penalty for large coefficients. X and Y represents respectively, the matrix of independent variables (also called predictors or features) and the vector of dependent variables. Cross-validation was used to select the optimal λ value, ensuring a balance between bias and variance [20].

$$\beta^{ridge} = \frac{X^T y}{X^T X + \lambda I} \quad (4)$$

Lasso regression, short for Least Absolute Shrinkage and Selection Operator, is a type of linear regression that incorporates regularization by adding a penalty term to the ordinary least squares (OLS) regression method. In contrast to ridge regression, which applies the L2 norm penalty, lasso regression employs the L1 norm penalty. This particular penalty promotes sparsity in coefficient estimates by penalizing the absolute magnitude of coefficients. As a result, lasso regression not only addresses multicollinearity like ridge regression but also performs variable selection by effectively setting some coefficients to zero, thus leading to a more interpretable and concise model. It's particularly useful when dealing with high-dimensional data where only a subset of predictors may be relevant. Lasso Regression, which performs both variable selection and regularization, was also applied. The Lasso regression coefficients βlasso are estimated by minimizing with equation 5 [21].

$$\beta^{lasso} = argmin\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad (5)$$

where λ penalizes the absolute size of the regression coefficients, driving some of them to zero and thus performing feature selection. The optimal λ was determined through cross-validation.

3. Results and Discussions

The first stage of the study is to graph CR and solar activity data. Figure 1 shows how the sunspot number and CR intensity correlate or interact over time. Changes in solar activity, reflected in the sunspot number, can influence the CR flux reaching the Earth. When solar activity is high (more sunspots), the Sun's magnetic field is stronger, which can deflect cosmic rays away from the inner solar system, leading to lower CR intensity on Earth.

The blue line representing the SSN shows the variation in the SSN over time. The red line representing CR shows the variation in the intensity of CRs reaching the Earth's atmosphere over time. CR are high-energy radiation originating from

outside the solar system. They can have various sources, including supernovae, black holes, and other cosmic events.

This period encompasses several solar cycles, as well as significant scientific and technological advancements. Solar cycles typically last around 11 years, during which the Sun undergoes a regular cycle of increasing and decreasing solar activity.

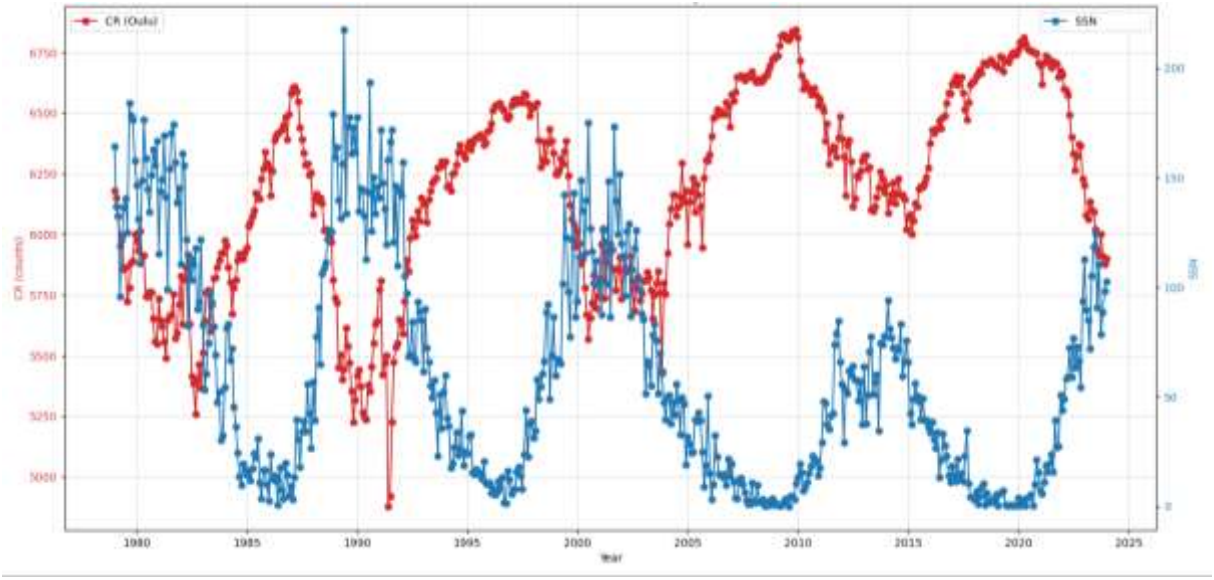

Figure 1. *Time Series of CR and SSN*

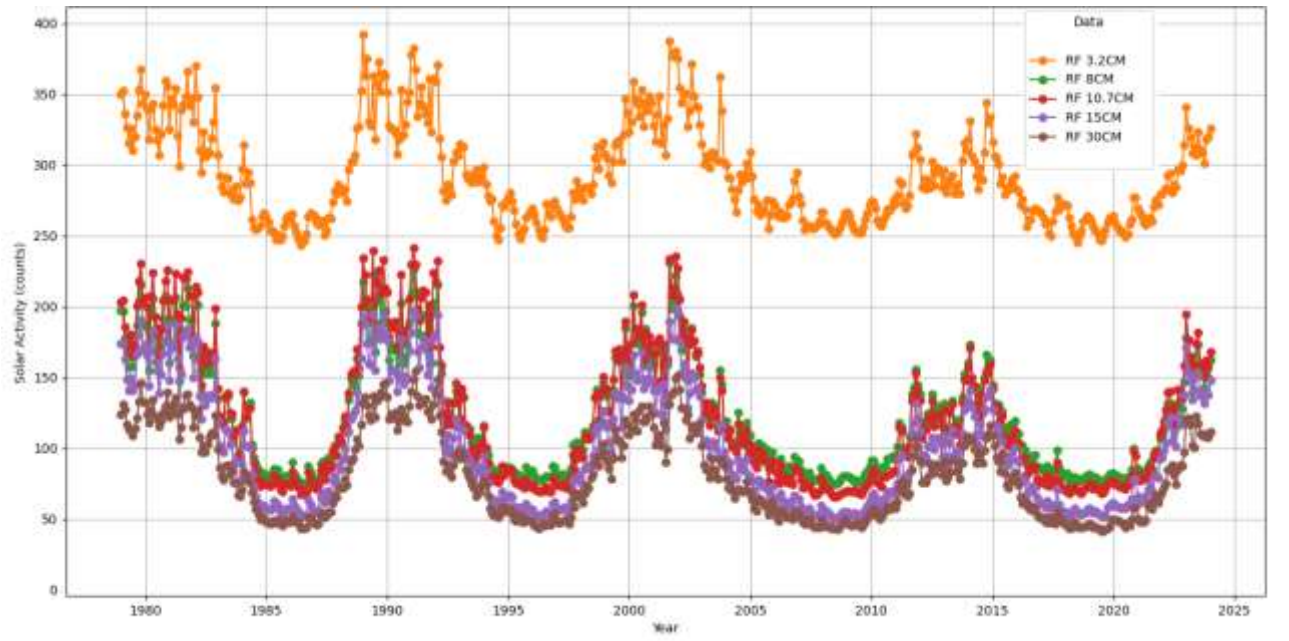

Figure 2. *Time Series of Solar Activity Parameters*

In this article, other parameters other than SSN were used among the solar activity parameters. These are Mg II Index, RF 3.2 cm, RF 8 cm, RF 10.7 cm, RF 15 cm, RF 30 cm and Lyman-alpha, respectively. These parameters also show the activity of the sun. In Figure 2, only the time graph of these parameters is drawn. The graph shows that all of these parameters are measured in direct proportion. More counts are obtained with the RF 3.2 cm parameter. Since the Lyman-alpha and Mg II Index parameters are set in the range of 0-1, it appears as a line on the graph. For this reason, these

data are included in Figure 3. Correlation analysis was performed to determine which of these parameters was more correlated with CR. Correlation analysis results are shown in Figure 4. According to this figure, the most opposite correlation with CR is seen at SSN and RF 30 cm with a value of -0.82. Mg II Index and RF 3.2 cm show a lower inverse correlation with values of -0.69 and -0.77 respectively. SSN and RF 3.2 cm parameters are very important in these studies, as a high negative correlation should be seen between CR and solar activity. However, other RF parameters can also be used with high accuracy.

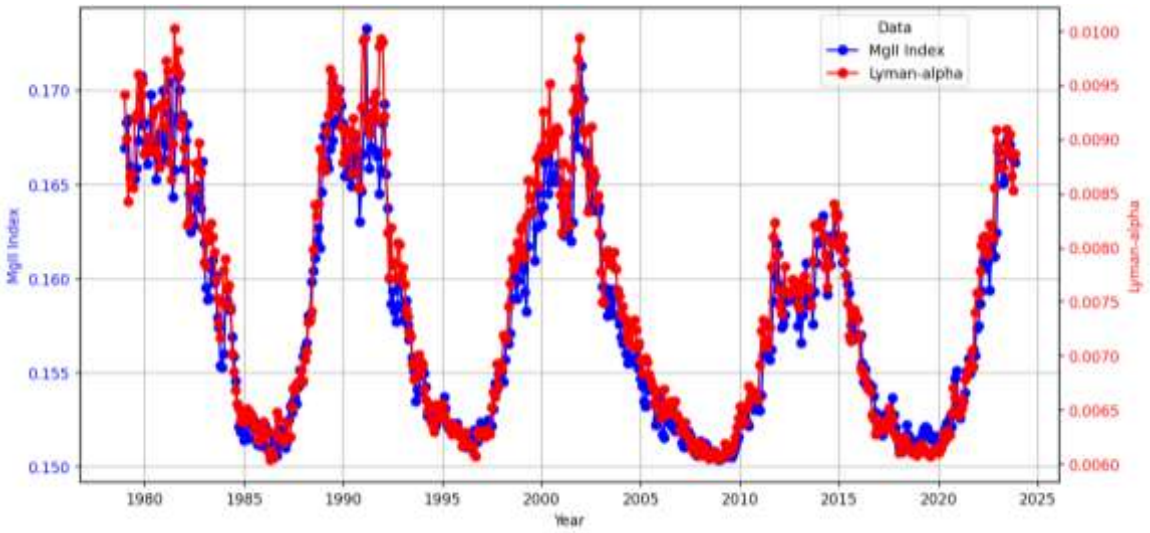

Figure 3. *Time Series of MgII Index and Lyman-alpha*

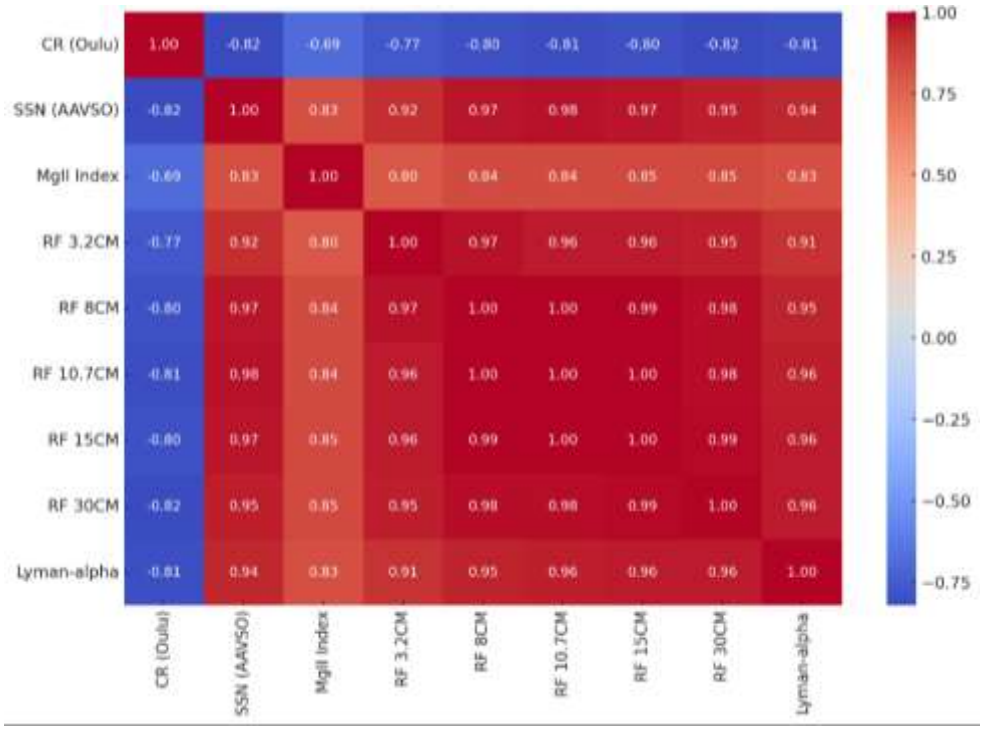

Figure 4. *Correlation Analysis of CR and Solar Activity Parameters*

The data we have has been evaluated with ML. In order to estimate CR values using the GBM model, the data set was first divided into training and test sets. Then, the GBM model was trained on the training set and the performance of the model was evaluated on the test set. When training the model,

CR values were used as the dependent variable. As independent variables, the features that correlate most strongly with CR will be selected. The following steps were followed in this process: The data was made available for modeling (separation into training and test sets, feature scaling if necessary), the GBM model was trained on the training set, and the performance of the model on the test set was evaluated and the results analyzed. The performance of predictions made using the GBM model includes the MSE and R^2 score calculated on the test set. The MSE value for the model was calculated as approximately 75458.82 and the R^2 score was approximately 0.51. These results show that the model can explain approximately 51% of the variance in the data. This suggests that we can make some useful predictions, but that the model is not perfect and perhaps other models or parameter settings that might perform better can be explored. The closer the R^2 score is to 1, the better the model explains the data, while the lower the MSE value, the better the model performs.

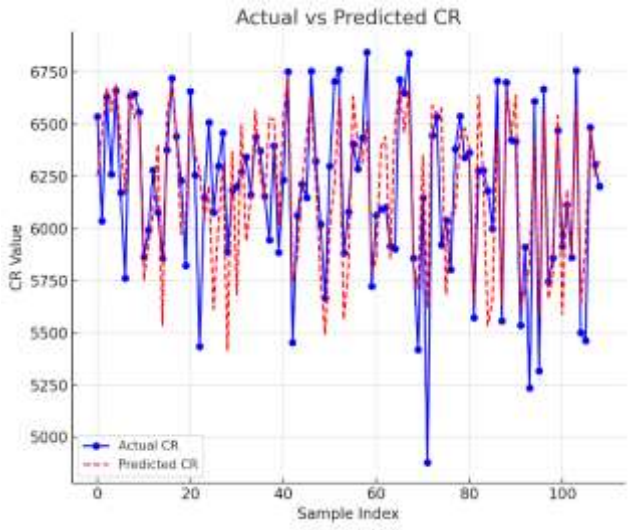

Figure 5. *Comparison of actual CR values and model-predicted CR values*

Experiments were made with different parameters to increase the performance of the model. When a GridSearchCV experiment was conducted with different parameter combinations, the best parameters for the GBM model were determined as 0.1 as the learning rate, 3 as the maximum depth (max_depth) and 100 as the number of trees (n_estimators). The R^2 score calculated from cross-validation with these parameters was approximately 0.71 and MSE was 70971.12. This result shows that our model can explain 71% of the variance in the dataset, which is an improvement over our first attempt above. With these parameters obtained, it can be said that the overall performance of the model is better. Using this model, the relationships between CRs and solar activity parameters can be better predicted. The final model was trained with these parameters. Figure 5 compares the actual CR values with the CR values predicted by the model. Blue dots represent actual values, and the red line represents the model's predictions. It is visualized how close the model's predictions are to the actual values.

With the same values, the results of the analysis with the ANN model show the performance of the model on the test set. For this model, the MSE is approximately 23287.40 and the R^2 score is approximately -151.28. These results show that the ANN model does not provide the expected performance on this data set. In particular, the negative R^2 score indicates that the model is far from explaining the variance in the data set and the predictions deviate significantly from the actual values. This indicates that the model has not yet been trained well enough or that the model configuration (e.g., number of layers, number of neurons, number of iterations) is not appropriate for the data set. In addition, a convergence warning was also received from the model. This indicates that the set maximum number of iterations (500) has been reached but the optimisation has not yet converged. This suggests that the model could be better trained with more iterations or may require a different configuration.

The results of the analysis with the Linear Regression model show the performance of the model on the test set. The MSE is approximately 62,163.04 and the R^2 score is approximately 0.59. This indicates a better performance compared to previous GBM and ANN models. The linear regression model is effective in capturing linear relationships in the data set. The MSE value indicates the magnitude of the model's errors and in this case, a lower value was obtained compared to the GBM and ANN models, indicating a better prediction performance.

By using regularisation techniques such as Ridge Regression or Lasso Regression, model overlearning can be avoided and overall performance can be improved. These techniques are particularly useful in data sets with a large number of features. Moving forward using regularisation methods makes the model more robust to overfitting and generally improves the prediction performance of the model. Finally, Ridge and Lasso regressions are used for this purpose. Both methods attempt to reduce overlearning by limiting the

complexity of the model, but they do so in different ways. Ridge Regression (L2 regularisation) adds the sum of the squares of the coefficients to the model as a penalty term. This penalises large coefficients and creates a smoother model. Lasso Regression (L1 regularisation) uses the sum of the absolute values of the coefficients as the penalty term. This can reduce some coefficients to zero, thus making automatic feature selection and making the model simpler.

Both models are trained and their performance is compared. This helps us to understand which regularisation method is more suitable for the dataset. Firstly, Ridge and Lasso models were trained. After evaluating the performance of the regression models, the MSE for the Ridge Regression was 66737.51 and the R^2 Score was 0.56. For Lasso Regression, MSE was 63018.65 and R2 Score was 0.59. These results indicate that the Lasso regression performs slightly better on your data set. The R^2 score of the Lasso model is 0.59, while the R^2 score of the Ridge model is 0.56. Also, the MSE of the Lasso model is lower than that of the Ridge model, indicating that the prediction errors are smaller on average. The automatic feature selection of the Lasso regression by reducing the coefficients of some features to zero is one of the reasons why this model is more effective in the data set. This is particularly useful in data sets with a large number of features or when some features carry little information. These results provide important information for model selection and regularisation strategy. It is possible to further optimise the model by trying different values for the regularisation parameter (alpha) or by applying other regularisation methods. When the effect of alpha values at different levels is analysed, it is observed that the performance of both Ridge and Lasso regression models varies depending on the alpha values. The results are shown in Table 1.

In Ridge Regression, the MSE increases with increasing alpha value, while the R^2 score decreases. This indicates that the model is less able to explain the variance in the data set. The best performance was observed when the alpha value was 0.01 and 0.1. In the Lasso Regression, similarly, as the alpha value increases, a decrease in the performance of Lasso is observed. The Lasso model also performs well when alpha is 0.01 and 0.1. This analysis shows that the value of the regularisation parameter (alpha) has a significant impact on model performance. For both models, lower alpha values provided better results in explaining the variance in the data set. This indicates that the model takes an appropriate approach in capturing the underlying structures of the dataset while avoiding over-learning. To optimise the model, one of these alpha values can be chosen, depending on the specific dataset and modelling objectives. It is also possible to systematically adjust the alpha value using cross-validation to further improve the model.

Table 1. *Changes in the performance of regression models depending on alpha values*

Alpha	MSE_{Ridge}	R^2_{Ridge}	MSE_{Lasso}	R^2_{Lasso}
0.01	62249.29	0.59	62558.99	0.59
0.1	62969.21	0.59	63018.65	0.59
1	66737.51	0.56	68375.51	0.55
10	70675.63	0.54	71621.27	0.53
100	70720.47	0.54	68204.91	0.55

The relationship between the predicted values and the actual values using the Lasso and Ridge regression models is shown in Figure 6. In both graphs, it is possible to see how close the predicted values are to the actual values. The red dots show the values predicted by the Lasso regression model and the actual values. The line is the identity line where perfect predictions should lie. For the Lasso regression, some of the dots are located close to the identity line, indicating that the model comes quite close to the true values in some predictions. Similarly, the blue dots show the predicted and actual values for Ridge Regression. For Ridge regression, the closeness of the points to the identity line indicates that the model is able to make appropriate predictions. While both models seem to capture some patterns in the dataset, there are deviations from perfect predictions. These deviations indicate that the model deviates from the true values in some cases. Such graphs are useful for visualising the performance of the model and assessing the accuracy of the predictions.

4. Conclusions

In this extensive study, we embarked on a journey to unravel the intricate dynamics between solar activity and cosmic ray intensity, employing a multifaceted approach encompassing data analysis, statistical methodologies, and advanced machine learning techniques. Through meticulous exploration of extensive datasets sourced from reputable repositories such as the LASP Interactive Solar IRradiance Datacenter (LISIRD) and the OULU neutron database, we endeavored to elucidate the underlying patterns, correlations, and implications of solar-cosmic interactions spanning decades. This investigation unveiled compelling insights into the profound influence of solar activity on cosmic ray modulation and its ramifications

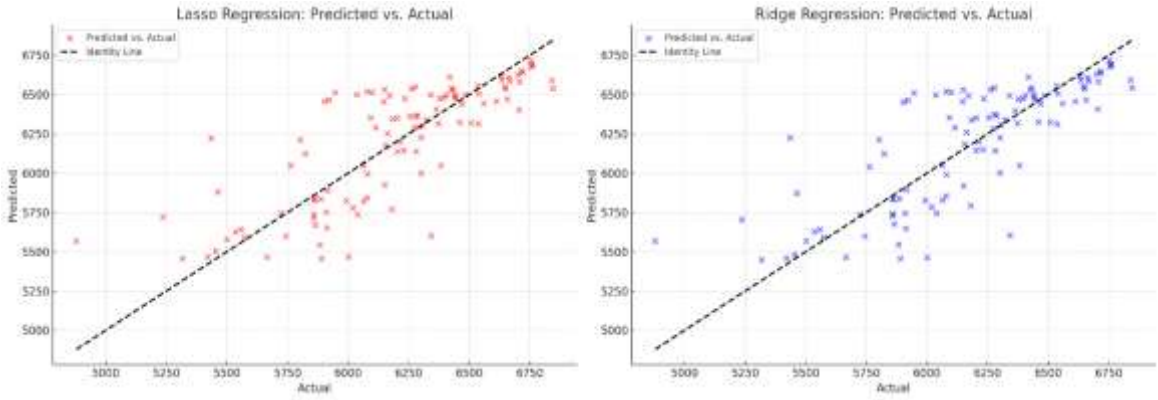

Figure 6. *Predicted and actual value using Lasso and Ridge regression models*

for space weather phenomena and Earth's environment. By meticulously examining solar activity parameters such as the sunspot number (SSN), Mg II Index, and various radio flux measurements (RF) across different wavelengths, we discerned robust correlations with cosmic ray intensity. Notably, our analysis revealed a strong inverse correlation between SSN and RF 30 cm, indicating the pivotal role of solar magnetic activity in shaping cosmic ray flux propagation through the heliosphere. Our discoveries unveil noteworthy associations among solar activity metrics like the SSN, Mg II Index, and diverse RF across varying wavelengths, and the intensity of cosmic rays. Particularly noteworthy is the robust negative correlation noted between SSN and RF 30 cm, quantified at -0.82, signifying the impact of solar activity on regulating the cosmic ray flow reaching the Earth. Through visualizations and statistical analyses, we showcased the intricate dance between solar dynamics and CR variability, underscoring the need for a holistic understanding of these phenomena for space weather forecasting and mitigation strategies.

Machine learning (ML) emerged as a potent tool in our quest to predict cosmic ray intensity based on solar activity parameters, offering unprecedented insights into the complex relationships embedded within the data. Models such as Gradient Boosting Machines (GBM) and Artificial Neural Networks (ANN) exhibited remarkable predictive capabilities, capturing nuanced patterns and nonlinear dependencies with remarkable accuracy. Through rigorous model training, validation, and optimization, we demonstrated the efficacy of machine learning in distilling actionable insights from vast datasets, paving the way for enhanced space weather forecasting and risk assessment methodologies. Furthermore, our exploration extended to the realm of regularization techniques, where Ridge and Lasso regression emerged as formidable allies in mitigating overfitting and enhancing prediction performance. By incorporating regularization strategies into our modeling framework, we attained greater resilience to data noise and improved generalization capabilities, thereby bolstering the robustness and reliability of our predictive models.

Quantitative analysis of our models revealed promising results, with GBM achieving a MSE of approximately 75458.82 and an R^2 score of approximately 0.51. Through parameter optimization using GridSearchCV, the performance of our models improved significantly, with an R^2 score of approximately 0.71 and an MSE of 70971.12. However, our experimentation with Artificial Neural Networks yielded less favorable results, with an MSE of approximately 23287.40 and a negative R^2 score, indicating suboptimal performance on the dataset. Incorporating Ridge and Lasso regression techniques further enhanced our models' performance, with Lasso regression exhibiting superior performance with an MSE of 63018.65 and an R^2 score of 0.59 compared to Ridge regression's MSE of 66737.51 and an R^2 score of 0.56. The analysis of regularization parameters demonstrated the significant impact of alpha values on model performance, underscoring the importance of parameter tuning in achieving optimal results.

The implications of our findings reverberate across diverse scientific domains, from astrophysics to atmospheric science, and hold profound implications for space exploration, satellite operations, and technological infrastructure. By elucidating the intricate interplay between solar activity and cosmic rays, our study lays the

groundwork for informed decision-making and proactive measures to mitigate the potential impacts of space weather events on society and technology. Looking ahead, continued research endeavors are imperative to further refine predictive models, explore novel analytical methodologies, and deepen our understanding of the solar-terrestrial relationship. By fostering interdisciplinary collaboration and leveraging cutting-edge technologies, we can unlock new frontiers in space weather forecasting, enhance our resilience to solar perturbations, and pave the way for a more secure and sustainable future in space exploration and technological advancement

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- **Acknowledgement:** The results presented in this document rely on data produced by the American Association of Variable Star Observers (AAVSO) (https://www.aavso.org/) and are available at https://www.aavso.org/solar. These data were accessed via the LASP Interactive Solar Irradiance Datacenter (LISIRD) (https://lasp.colorado.edu/lisird/). In addition, Cosmic Ray data were obtained from Oulu University, Cosmic Ray station (https://cosmicrays.oulu.fi/).
- **Author contributions:** The author declares that they have equal right on this paper.
- **Funding information:** The author declares that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Schrijver, C. J., Bagenal, F., & Sojka, J. J. (Eds.). (2016). Heliophysics: Active stars, their astrospheres, and impacts on planetary environments. Cambridge University Press.

[2] Hachaj, T., Bibrzycki, Ł., & Piekarczyk, M. (2023). Fast training data generation for machine learning analysis of cosmic ray showers. *IEEE Access*, 11, 7410-7419.

[3] Malinović-Milićević, S., Radovanović, M. M., Radenković, S. D., Vyklyuk, Y., Milovanović, B., Milanović Pešić, A., ... & Gajić, M. (2023). Application of solar activity time series in machine learning predictive modeling of precipitation-induced floods. *Mathematics*, 11(4), 795.

[4] Kumar, P., Pal, M., Rani, A., Mishra, A. P., & Singh, S. (2022). Modulation of Cosmic Ray with Solar activities During Solar Cycles 19-24 to forecast *Solar Cycle* 25.

[5] Verbanac, G., Vršnak, B., Temmer, M., Mandea, M., & Korte, M. (2010). Four decades of geomagnetic and solar activity: 1960–2001. *Journal of atmospheric and solar-terrestrial physics*, 72(7-8), 607-616.

[6] Drury, L. O. C. (2012). Origin of cosmic rays. *Astroparticle Physics*, 39, 52-60.

[7] Bazilevskaya, G. A., Cliver, E. W., Kovaltsov, G. A., Ling, A. G., Shea, M. A., Smart, D. F., & Usoskin, I. G. (2014). Solar cycle in the heliosphere and cosmic rays. *Space Science Reviews*, 186, 409-435.

[8] Potgieter, M. S. (2013). Solar modulation of cosmic rays. *Living Reviews in Solar Physics*, 10, 1-66.

[9] Mohamed, A. E. (2017). Comparative study of four supervised machine learning techniques for classification. *International Journal of Applied*, 7(2), 1-15.

[10] Patel, V. R., & Mehta, R. G. (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. *International Journal of Computer Science Issues* (IJCSI), 8(5), 331.

[11] Hatfield, P. W., Gaffney, J. A., Anderson, G. J., Ali, S., Antonelli, L., Başeğmez du Pree, S., ... & Williams, B. (2021). The data-driven future of high-energy-density physics. *Nature*, 593(7859), 351-361.

[12] Laboratory for Atmospheric and Space Physics. (2005). LASP Interactive Solar Irradiance Datacenter. *Laboratory for Atmospheric and Space Physics*. https://doi.org/10.25980/L27Z-XD34

[13] Kananen, H., P.J. Tanskanen, L.C. Gentile, M.A. Shea and D.F. Smart, A quarter of a century of relativistic solar cosmic ray events recorded by the Oulu neutron monitor, *Proc. 22nd ICRC*, 3, 145-148, 1991.

[14] Jebli, I., Belouadha, F. Z., Kabbaj, M. I., & Tilioua, A. (2021). Prediction of solar energy guided by pearson correlation using machine learning. *Energy*, 224, 120109.

[15] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peerj computer science*, 7, e623.

[16] Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308-324.

[17] Kartini, D., Nugrahadi, D. T., & Farmadi, A. (2021, September). Hyperparameter tuning using GridsearchCV on the comparison of the activation function of the ELM method to the classification of pneumonia in toddlers. In 2021 *4th International Conference of Computer and Informatics Engineering (IC2IE)* (pp. 390-395). IEEE.

[18] Alaloul, W. S., & Qureshi, A. H. (2020). Data processing using artificial neural networks. Dynamic data assimilation-beating the uncertainties.

[19] Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(2), 140-147.

[20] Kim, H., & Jung, H. Y. (2020). Ridge fuzzy regression modelling for solving multicollinearity. *Mathematics*, 8(9), 1572.

[21] Ahrens, A., Hansen, C. B., & Schaffer, M. E. (2020). lassopack: Model selection and prediction with regularized regression in Stata. *The Stata Journal*, 20(1), 176-235.