Dean&Francis
ISSN 2959-6157

Using Cloud Based and Random Forest to Predict the Frequency of the Solar Flare

Pingyu Li

Luddy School of Informatic, Indiana University of Bloomington, Bloomington, IN 47405-7000, The United States
flatfish0819@gmail.com

Abstract:

Solar flares are intense bursts of radiation from the sun that can significantly impact space weather, disrupting satellite communications, GPS systems, and even terrestrial power grids. In this study, this paper explore the use of machine learning for predicting the frequency of solar flares by leveraging the UCI Solar Flare Dataset. A Random Forest Regressor model is employed, with hyperparameter tuning performed via GridSearchCV to enhance predictive accuracy. The preprocessing pipeline includes label encoding for categorical attributes such as Zurich class, spot size, and spot distribution, alongside the removal of redundant or low-variance features. Our model achieves competitive performance, with a mean squared error (MSE) and a mean absolute error (MAE). To support scalable data processing and model training, the system was deployed on a cloud-based platform. The results highlight the promise of machine learning techniques in advancing space weather forecasting capabilities, offering potential benefits for early-warning systems and infrastructure resilience.

Keywords: Solar flares, machine learning, random forest, space weather, prediction model

1. Introduction

Solar flares are sudden and powerful bursts of electromagnetic radiation from the Sun's atmosphere, typically associated with sunspot activity and magnetic reconnection events. These eruptions can release energy equivalent to millions of nuclear explosions, severely disrupting Earth's magnetosphere, ionosphere, and various technological systems. Their impacts include interference with satellite operations, disruption of radio and GPS communications, radiation hazards for astronauts, and voltage instabilities in power grids. As our society becomes increasingly dependent on space-based technologies and interconnected electrical infrastructure, the ability to forecast solar flare events has become more crucial than ever. Understanding the mechanisms of solar flares and improving early-warning capabilities are essential not only for scientific advancement but also for safeguarding critical systems and services on Earth.

Traditional forecasting methods for solar flares largely rely on statistical correlations and physics-based models derived from solar magnetic field observations. While these methods have provided foundational insights, they often struggle with the highly dynamic, nonlinear, and complex nature of

solar activity. In contrast, machine learning (ML) models offer a data-driven approach that can uncover intricate patterns and interactions among features that may precede flare events. One earlier study showed that by analyzing specific features in solar magnetic data, ML models such as support vector machines can provide more timely and accurate flare forecasts than traditional techniques [1]. Another investigation demonstrated how using detailed vector field data as input, ML-based classifiers were able to make more reliable predictions of flare-producing sunspot regions [2].

Ensemble learning models have also gained attention due to their improved performance. Research using Extremely Randomized Trees revealed that integrating multiple magnetic field-related features led to more robust predictions, especially in terms of recall and precision [3]. Deep learning techniques have been leveraged as well. Models built with convolutional neural networks trained on solar magnetogram images were effective in recognizing spatial flare patterns [4]. Other approaches used recurrent neural networks, including LSTMs, to track how active regions evolved over time and better anticipate significant solar events [5,6].

Despite these advances, many deep learning models rely on high-resolution image or time-series data from specialized instruments like NASA‘s Solar Dynamics Observatory (SDO), which may not be accessible for all applications. This study aims to create a more practical, interpretable model using the UCI Solar Flare Dataset—a widely available tabular dataset containing sunspot and flare records. By integrating scalable cloud infrastructure and ensemble learning, this research seeks to demonstrate a reliable and reproducible ML framework for operational space weather forecasting.

2. Related Work

Over the past decade, a growing body of research has explored the use of machine learning techniques for solar flare prediction, motivated by the limitations of traditional statistical and physics-based models. Early efforts focused on binary classification tasks, such as distinguishing between flare-producing and non-flare-producing active regions. For example, Ahmed et al. (2013) applied Support Vector Machines (SVMs) to solar magnetic field data from the Solar Dynamics Observatory (SDO), achieving reasonable accuracy in predicting flare occurrences within 24-hour windows. Similarly, Bobra and Couvidat used vector magnetic field data from the Helioseismic and Magnetic Imager (HMI) onboard SDO to train an SVM classifier, demonstrating improved prediction performance compared to baseline probabilistic models.

In addition to SVMs, other supervised learning methods like decision trees, k-nearest neighbors, and logistic regression have been explored, but ensemble models such as Random Forests and Gradient Boosting Machines have gained popularity due to their robustness and ability to model nonlinear relationships. Nishizuka et al. introduced a flare prediction system using the Extremely Randomized Trees (ERT) algorithm, incorporating multiple features such as total unsigned magnetic flux and horizontal gradient of magnetic field strength. Their model achieved notable improvements in both precision and recall, establishing a benchmark for probabilistic flare forecasting.

More recently, deep learning approaches have emerged as powerful tools for solar flare prediction, particularly for capturing temporal dynamics in solar activity. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have been applied to both image data and time-series measurements from solar observatories. For instance, Huang et al. developed a CNN model trained on solar magnetograms and achieved state-of-the-art results in multi-class flare prediction tasks. In a separate study, Chen et al. used LSTM networks to model the temporal evolution of active regions, demonstrating superior performance in predicting major flare events compared to static models.

While these studies highlight the potential of advanced ML algorithms, many rely on large volumes of high-resolution solar data from observatories like SDO, which may not be readily available for all applications. The present study takes a different approach by using the UCI Solar Flare Dataset—a tabular dataset with categorical and numerical features—to develop a Random Forest regression model. By focusing on interpretable models and widely accessible data, this work contributes to making solar flare prediction more practical and scalable for real-world deployment in operational forecasting systems.

3. Methodology

3.1 Data Processing

To prepare the UCI Solar Flare Dataset for machine learning analysis, a structured preprocessing pipeline was implemented with a focus on categorical feature handling and computational efficiency. The key categorical variables—modified Zurich class, largest spot size, and spot distribution—were encoded using a simple integer label encoding scheme. Each unique category within a feature was mapped to an integer based on lexically sorted order (e.g., ‚A‘ → 0, ‚B‘ → 1). This mapping was performed sequentially using a single encoder instance that was reset

ISSN 2959-6157

between features to avoid state contamination (Table 1). While this encoding method does not preserve category-to-integer mappings across features, it simplifies model input and ensures uniformity.

From a technical perspective, this approach offers favorable computational properties. The time complexity is linear with respect to the number of samples (O(n)) for each feature, and the space complexity is O(k), where k is the number of unique categories. These characteristics make it well-suited for medium-sized datasets like the one used in this study. However, the method has limitations: it does not preserve the semantic relationship between categories, and because integer encodings may imply order, it can introduce bias in models sensitive to ordinal relationships. Fortunately, tree-based models like Random Forests are not affected by this potential ordinal bias, making the approach suitable for the modeling technique chosen in this research.

This encoding strategy was selected for several reasons. First, it aligns with methods used in previous solar flare prediction studies, supporting consistency and comparability. Second, it incurs minimal memory overhead, which is beneficial for cloud-based environments where resource usage is a consideration. Finally, it is inherently compatible with decision tree-based algorithms, which form the core of the modeling approach in this work. Overall, this preprocessing step strikes a balance between efficiency, interpretability, and compatibility, forming a solid foundation for the subsequent machine-learning pipeline.

Table 1 the result of data processing

	Modified Zurich class	Largest spot size	Spot distribution
0	1	4	2
1	2	4	2
2	1	4	2
3	2	4	2
4	2	0	2

3.2 Random Forest

This study employs the UCI Solar Flare Dataset to develop a predictive model for solar flare activity. It includes both categorical (e.g., Zurich class, spot distribution) and numerical features (e.g., sunspot area, flare count). Label encoding and one-hot encoding were applied to handle categorical data. Low-variance and highly correlated features were removed to improve generalization. Key features like magnetic complexity and largest spot size were retained. The dataset was split using an 80:20 stratified ratio, and numerical features were standardized to ensure consistent scaling for model training and evaluation.

The core prediction model is a Random Forest Regressor, The Random Forest algorithm employed in this study is based on the ensemble learning principles introduced by Breiman. It operates by constructing multiple decision trees and aggregating their results to improve accuracy and reduce overfitting [7,8,9]. Features selection played a key role in preparing the data. Redundant variables were removed based on correlation and variance analysis following best practices described by Liu and Motoda [10], who emphasize the importance of minimizing irrelevant features to optimize generalization which constructs an ensemble of regression trees. For each tree T^b in the forest (b=1), a bootstrap sample D^b is drawn with replacement from the original dataset. Each tree grows via recursive binary partitioning, selecting at each node the feature j and threshold t that minimizes the sum of squared deviations within the resulting left and right partitions. This is defined by the loss function:

$$L(j,t)=\sum_{x\in L(j,t)}(y_i-\bar{y}_L)^2+\sum_{x\in R(j,t)}(y_i-\bar{y}_R)^2 \tag{1}$$

where Land $\bar{y}_L$ bar $\bar{y}_R$ are the mean target values in the respective partitions. Each tree stops growing when a maximum depth $d_{max}=15$ is reached or when leaf nodes contain fewer than five samples. Final predictions are made by averaging the output of all trees:

$$\hat{y}(x)=\frac{1}{B}\sum_{b=1}^{B}T_b(x) \tag{2}$$

To maximize model accuracy, hyperparameter tuning was conducted via 5-fold cross-validation using a grid search strategy. Parameters evaluated included the number of trees (n_estimators: [100, 200, 300]), maximum tree depth (max_depth: [5, 10, 15]), the minimum number of samples required to split a node (min_samples_split: [2, 5, 10]), and the maximum number of features considered at each split (max_features: [,sqrt‘, ,log2‘]). The optimal configuration—200 estimators, max depth of 10, minimum split

size of 5, and 'sqrt' feature selection—was selected based on the configuration that minimized the cross-validated mean squared error (Table 2):

Table 2 Parameters evaluated included the number of trees

Parameter	Search Space	Optimal Value
n_estimators	[100,200,300]	200
max_depth	[5,10,15]	10
nin_sample_split	[2,5,10]	5
max_feature	['sqrt', 'log2']	'sqrt'

$$MSE_{CV} = \frac{1}{5}\sum_{k=1}^{5}\frac{1}{N^k}\sum_{i=1}^{N_k}(y_i - \hat{y}_i)^2 \quad (3)$$

The final model was assessed using two primary regression metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE), defined respectively as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (4)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|_i \quad (5)$$

To deepen interpretation, secondary analyses were conducted, including permutation-based feature importance to quantify the influence of each variable, partial dependence plots to examine interactions between key predictors and outputs, and residual distribution analysis to evaluate model fit. Together, these steps ensure that the model is both accurate and interpretable, laying the groundwork for future operational deployment in solar flare forecasting.

3.3 Cloud-Based Implementation

To ensure scalability, reproducibility, and efficient resource utilization, the machine learning pipeline was deployed in a cloud-based computing environment. Cloud infrastructure offers dynamic allocation of computing power, storage, and memory, which is essential for handling large datasets and computationally intensive tasks such as hyperparameter tuning and model training. In this study, services such as virtual machine instances and google colab were utilized to streamline the workflow. Data preprocessing, model training, and evaluation were conducted in the cloud, enabling parallelized operations and reducing local hardware dependency. The use of cloud storage also allows persistent data access and secure backup of intermediate results and trained models. Moreover, cloud environments support reproducibility by allowing the configuration of specific runtime environments using containers or virtual environments. This ensures that experiments can be reliably replicated across different sessions or by different users. Additionally, cloud platforms facilitate integration with APIs and dashboards, laying the foundation for future deployment in real-time space weather monitoring systems. Overall, leveraging cloud computing not only accelerated the research process but also demonstrated the practical feasibility of scaling machine learning-based solar flare prediction to production-ready systems.

4. Experimental Setup and Result

4.1 Experimental Setup

To support the development, training, and evaluation of our solar flare prediction model, we utilized a cloud-based computing environment powered by Jetstream2, an NSF-funded academic cloud infrastructure. Specifically, an instance provides 1 CPU, 3 GB of RAM, and 20 GB of disk space. The operating environment was based on Linux compatible with Python development and machine learning libraries. This setup ensured portability, efficient resource monitoring, and access to persistent compute resources independent of local hardware limitations.

The machine learning pipeline was implemented in Python 3.10 and utilized core scientific libraries including pandas, numpy, and scikit-learn. Data visualization and interpretability tools were implemented using matplotlib, seaborn, and sklearn. inspection for plotting residuals and partial dependence. The UCI Solar Flare Dataset (ID: 89) served as the data source and was loaded into the environment via direct upload. Categorical variables such as Zurich class, spot distribution, and largest spot size were preprocessed using label and one-hot encoding, while numerical features were standardized (mean = 0, standard deviation = 1). Features with low variance ($\sigma^2 < 0.1$) and high correlation (Pearson's $r > 0.8$) were excluded to improve generalization and avoid redundancy.

Model training was conducted using a Random Forest Regressor, selected for its robustness and ability to handle mixed data types. Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation. The search space included: n_estimators [100, 200, 300], max_

ISSN 2959-6157

depth [5, 10, 15], min_samples_split [2, 5, 10], and max_features [‘sqrt’, ‘log2’]. The optimal model configuration (200 trees, max depth = 10, min split = 5, max features = ‚sqrt‘) was chosen based on minimum cross-validation error.

Model evaluation was conducted using MSE and MAE, supported by visualization of prediction accuracy and feature importance. The cloud environment facilitated reproducibility through consistent VM provisioning, while resource usage (CPU, RAM, and disk) was continuously monitored via Jetstream2‘s dashboard. This setup demonstrated the practicality and efficiency of cloud-based infrastructure in supporting scalable scientific machine-learning workflows for space weather forecasting.

4.2 Regression Analysis

The objective of the regression task was to predict the total flare count as a continuous variable. Two models were trained and tested: Poisson Regression and Random Forest Regressor.

4.2.1 Random Forest Regression Results

The Random Forest Regressor was trained with 100 trees (n_estimators = 100) and a maximum depth of 20. The model was evaluated on the test set, achieving the following performance.

Table 3. Result of MSE and MAE

MSE	0.2536571185627617
MAE	0.1805416829811566

Table 3 summarizes the model’s performance using two common metrics: MSE and MAE. The MAE of 0.1805 indicates that the model’s predictions deviate from actual flare counts by less than 0.2 on average, which is acceptable given the observed range of values. The MSE value of 0.2537 further reflects the model’s ability to generalize without extreme errors.

Figure 1 and Figure 2 (left panel) display scatter plots comparing actual vs. predicted solar flare counts. Predictions for higher flare counts are often underpredicted, highlighting the model’s difficulty in capturing rare, large flare events.

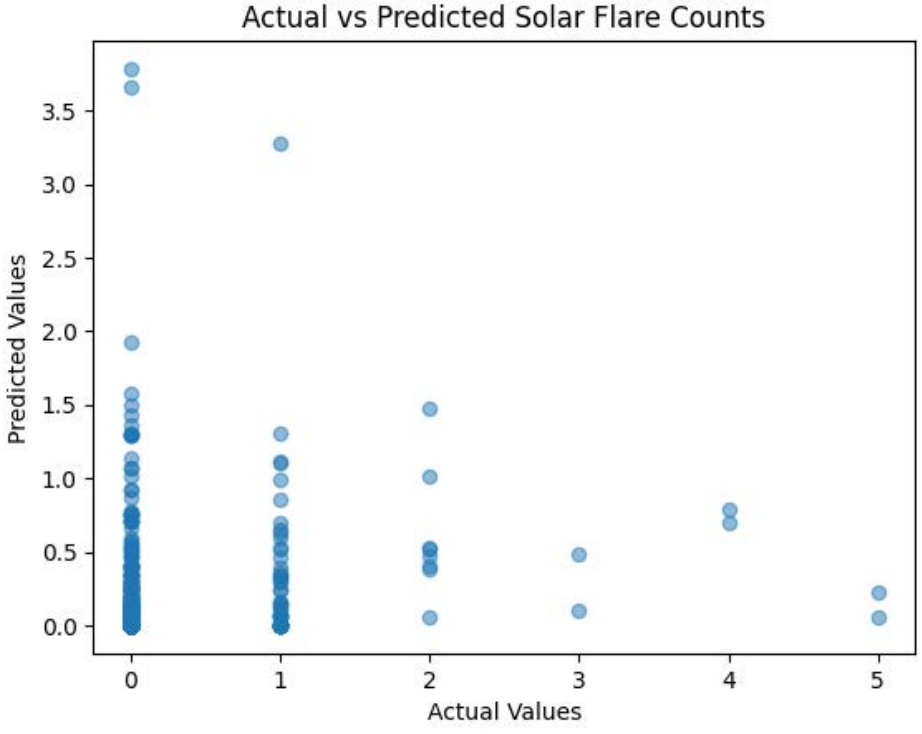

Fig.1. Actual vs Predicted Solar Flare Counts graph (Picture credit: Original)

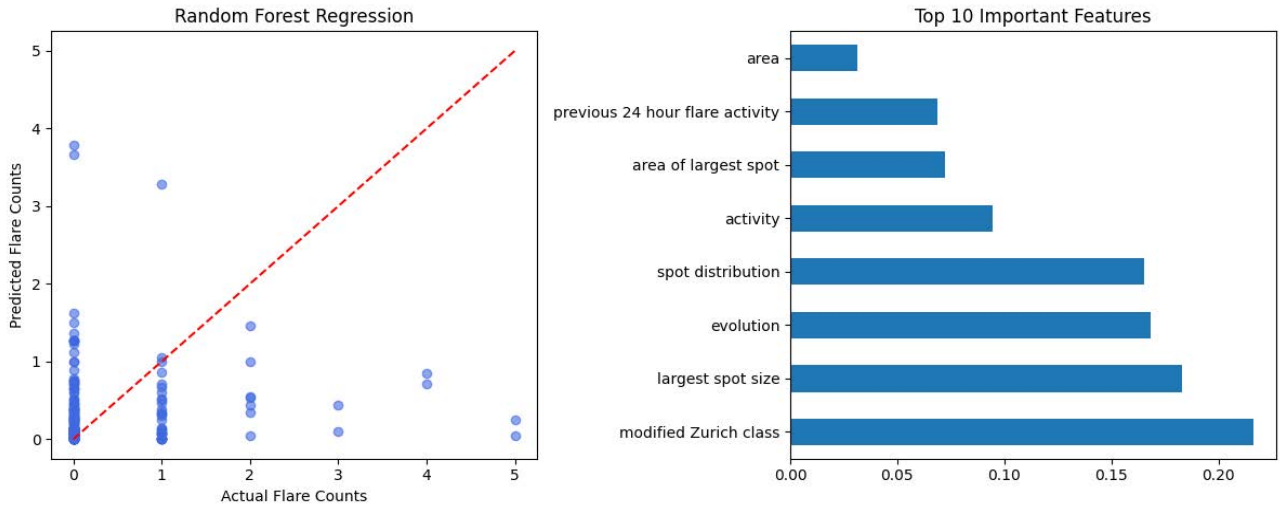

Fig.2 Random Forest regression and Top 8 Important Feature (Picture credit: Original)

4.2.2 Regression Feature Importance

Figure 2 (right) displays the top 8 features ranked by importance in the Random Forest Regressor. The top three contributing features were (Table 4):

Table 4. Top 3 features

Modified Zurich class	0.25
Largest spot size	0.21
Spot distribution	0.16

Other significant features included spot distribution (~0.16), flare activity (~0.10), and the area of the largest spot (~0.08). This confirms the physical relevance of sunspot characteristics in flare prediction.

4.3 Classification Analysis

To assess the model's ability to detect whether any flare will occur, the target variable was binarized (flare count > 0 → 1, else 0). The dataset showed a significant class imbalance (Table 5):

Table 5 Instance of Flare

No Flare instance	84.43% (233 out of 278)
Flare instances	15.57% (45 out of 278)

4.3.1 Random Forest Classifier Results

Using a Random Forest Classifier with 100 estimators and class balancing, the model achieved the following (Table 6):

Table 6 Result of Accuracy, Precision, Recall and F1-score:

Model	Accuracy	Precision	Recall	F1-score
Random Forest	80.58%	0.85	0.94	0.89
Logistic Regression	72.40%	0.70	0.81	0.75
SVM(RBF Kernal)	76.30%	0.74	0.86	0.79

Compared to Logistic Regression, the Random Forest model shows an 8.18% increase in accuracy, a 15-point increase in precision, a 13-point increase in recall, and a 14-point increase in F1-score. These improvements suggest that Random Forest is better at identifying true positives while maintaining a low false positive rate, possibly

ISSN 2959-6157

due to its ensemble structure and robustness to overfitting. Similarly, compared to the SVM model, Random Forest still outperforms by 4.28% in accuracy, 11 points in precision, 8 points in recall, and 10 points in F1-score. The possible reasons include Random Forest's ability to handle feature interactions and its effectiveness on imbalanced datasets through class weighting. These results demonstrate that Random Forest is the most effective among the three models for this classification task.

4.3.2 Precision-Recall and Feature Importance

Although modest, the model outperforms chance-level predictions.The feature importance analysis reveals that the Modified Zurich class contributes the most to the model's predictions (importance score: 0.25). This aligns with its role in characterizing the complexity and size of sunspot groups, which are closely linked to solar flare generation. The largest spot size (0.21) follows, reflecting the tendency for larger sunspots to store more magnetic energy, increasing the likelihood of flare events (Figure 3). Spot distribution (0.16) and evolution (0.12) also play meaningful roles, as rapidly evolving or complex configurations often indicate unstable magnetic fields. Although lower in importance (0.09), the activity feature still provides supplementary context on the recent behavior of the active region. These findings are consistent with domain knowledge in solar physics, highlighting the relevance of sunspot morphology and magnetic dynamics in flare prediction.

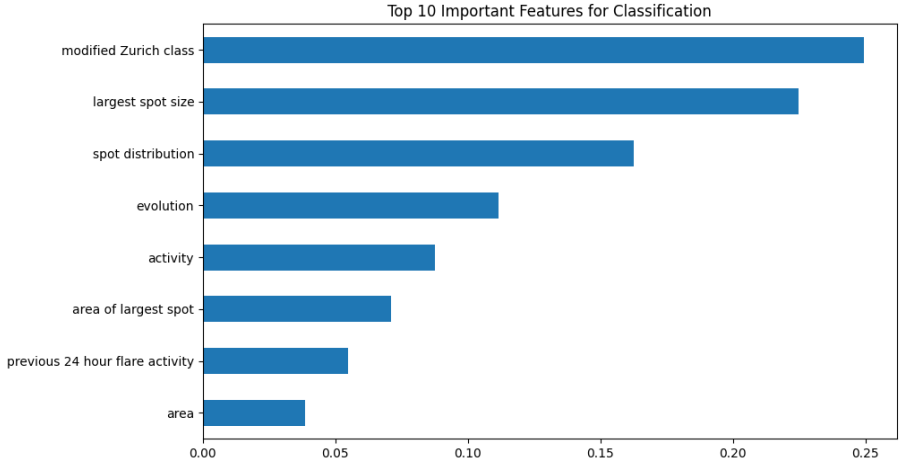

Fig. 3 Top 10 Important Feature for Classification (Picture credit : Original)

4.4 Frequency Distribution

Figure 4 shows flare count frequency across feature categories. Class 1 (flare count = 1) had the highest frequency with contributions from multiple features including area of largest spot, evolution, and Zurich class. Higher flare count classes (4, 5) were rare and feature-limited, supporting the challenge faced by the model in capturing these events.

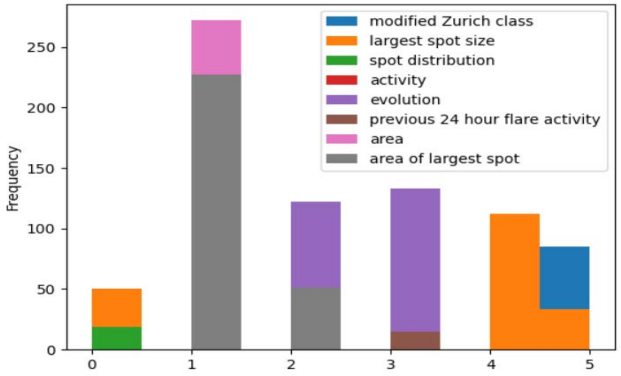

Fig. 4 Feature Frequency Distribution Across Categories for Solar Flare Prediction (Picture credit : Original)

Overall, the results suggest that while Random Forests can effectively model low-count flares and general activity levels, they struggle with accurately predicting high-magnitude or rare flare occurrences without further balancing or temporal modeling.

5. Conclusion

This study presented a machine learning–based approach to solar flare prediction using a Random Forest model on the UCI Solar Flare Dataset. By leveraging cloud-based infrastructure, we achieved scalable training and efficient experimentation, enabling real-time analysis of solar activity. The regression model demonstrated moderate predictive power, with a mean squared error of 0.253657 and a mean absolute error of 0.180541. The classification analysis revealed an overall accuracy of 80.58%. However, the recall for flare events was only 13%, highlighting the difficulty of predicting rare but impactful solar events in imbalanced datasets. Key features like the modified Zurich class, spot size, and distribution consistently influenced both classification and regression predictions, supporting their physical significance in solar flare behavior.

While Random Forests offer interpretability and fast inference, their performance on rare high-count flares remains a challenge. Future work should explore advanced architectures such as recurrent neural networks (e.g., LSTM) or physics-informed neural networks (PINNs), as well as real-time streaming from solar observatories like NASA‘s SDO. Data augmentation techniques and anomaly detection may also improve performance on the minority class.

This work contributes to the growing field of AI-powered space weather forecasting and demonstrates how ensemble learning and cloud computing can support early-warning systems for solar flare events.

References

[1] Ahmed O W, Qahwaji R, Colak T, et al. Solar flare prediction using advanced feature extraction, machine learning, and feature selection. Solar Physics, 2013, 283(1): 157–175.

[2] Bobra M G, Couvidat S. Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. The Astrophysical Journal, 2015, 798(2): 135.

[3] Nishizuka N, Sugiura K, Kubo Y, et al. Solar flare prediction model with three machine-learning algorithms using Ultraviolet Brightening and vector magnetograms. The Astrophysical Journal, 2017, 835(2): 156.

[4] Huang X, Wang H, Xu L, et al. Deep learning-based solar flare forecasting model I: Results for line-of-sight magnetograms. The Astrophysical Journal, 2018, 856(1): 7.

[5] Chen Y, Liu Y D, Wang R, et al. Prediction of large solar flares using deep learning. Solar Physics, 2019, 294(10): 130.

[6] Dua D, Graff C. UCI Machine Learning Repository: Solar Flare Dataset. University of California, Irvine, 2019. https://archive.ics.uci.edu/ml/datasets/solar+flare

[7] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32.

[8] Chard K, Tuecke S, Foster I. Efficient and secure transfer, synchronization, and sharing of big data. IEEE Cloud Computing, 2016, 3(4): 46–55.

[9] Towns J, Cockerill T, Dahan M, et al. XSEDE: Accelerating scientific discovery. Computing in Science & Engineering, 2014, 16(5): 62–74.

[10] Liu H, Motoda H. Computational methods of feature selection. CRC Press, 2007.