JURNAL UPI

Journal of Computers for Society

Journal homepage: https://ejournal.upi.edu/index.php/JCS

Journal of Computers for Society

Predicting Solar Flares Using Data Products Vector Magnetic SDO/HMI dan Random Ferns

Rooseno Rahman Dewanto, Lala Septem Riza, Judhistira Aria Utama*

Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia
*Correspondence: E-mail: rooseno.rahman.d@student.upi.edu

ABSTRACT

Solar flares (SFs) are the most powerful bursts of energy in the solar system that often have a bad effect on space weather. Until now, the cause of its appearance is not known for sure. Nevertheless, SFs are known to have magnetic properties attached to them. Therefore, understanding the configuration of the magnetic field on the sun plays an important role in SFs prediction efforts. Using SFs flux data recorded by X-ray Sensors on the Geostationary Operational Environmental Satellite (GOES) which is mapped with 13 parameters of the magnetic vector data of the solar photosphere layer recorded by the Helioseismic and Magnetic Imager (HMI) at the Solar Dynamic Observatory (SDO) and the Machine Learning (ML) Random Ferns (RFe) algorithm, This study tries to predict the emergence of multiclass SFs (B, C, M, and X) along with binary SFs (BC and MX). This study uses data from May 1, 2010 to May 10, 2020, with a total of 30 classes X, 443 classes M, 1032 classes C, 751 classes B, 473 classes MX, and 1783 classes BC. This study also applies the oversampling method to handle the imbalanced nature of the data on SFs data. Overall, it can be seen that predicting the occurrence of SFs using RFe is a valid effort. The highest average scores achieved by this study for sensitivity/recall, precision, and True Skill Statistics (TSS) in multiclass SFs were 74.4%, 50.3%, and 58.7%, respectively; and in binary SFs are 87.7%, 77.7%, and 72.8%.

ARTICLE INFO

Article History:
Submitted/Received 20 May 2023
First Revised 18 Jun 2023
Accepted 12 Aug 2023
First Available Online 13 Aug 2023
Publication Date 15 Sep 2023

Keyword:
Geostationary operational,
Environmental satellite,
Helioseismic and magnetic imager,
Oversampling,
Predictions,
Random ferns,
Solar dynamic observatory,
Solar flares,
Vector magnetic,
X-ray sensors.

1. INTRODUCTION

The sun plays an important role for life on earth. The energy it emits is the main source for plants to photosynthesize (Ruban, 2015), become raw materials for electricity conversion through solar panels (Grätzel, 2007), and become a source of light for animal and human vision. In the early days of human civilization, the sun even received special reverence with the manifestation of structures such as Newgrange in Ireland, Stonehenge in England, and Chankillo in Peru (Ghezzi and Ruggles, 2007). Various efforts to observe the sun and related phenomena continue to be carried out, including one of them is Solar Flares (SFs).

SFs are the most powerful eruption phenomenon in the solar system because they can release as much as 1032 ergs of energy (Emslie *et al*., 2012). This energy is ten million times greater than the energy released by volcanic eruptions . The radiation it emits is almost across the entire electromagnetic spectrum, ranging from 0.002 Å (2 ×10-11 cm, 6.1 MeV) to more than 10 km (106 cm, 30 kHz). This SFs phenomenon is also known to have a significant negative impact on the magnetosphere, atmosphere, and the environment inside the earth, such as causing electronic damage to spacecraft, radio communication interference, power network breaks, damage to transformers, blocking radar operations, and even damaging submarine cable networks. A study from the National Research Council of the United States in 2008 even predicted that if the SFs (Carrington Event) event like in 1859 hit the earth again, the losses that would be experienced by global civilization would be 2 trillion US dollars.

The magnitude of the impact that SFs can cause has prompted many researchers to find out about how SFs are formed and what factors underlie their emergence. Although the mechanism of energy release in SFs is not yet fully understood, SFs are known to have magnetic properties (Priest and Forbes, 2002). Therefore, the study of the configuration of the magnetic field in the sun's atmosphere is very important to understand and predict SFs (Bobra and Couvidat, 2015). Until now, one of the instruments that continuously observes and records the sun's magnetic activity is the Helioseismic and Magnetic Imager (HMI) on the Solar Dynamic Observatory (SDO). Launched on February 11, 2010, SDO then orbited on a Geosynchronous Orbit (GSO) with an inclination angle of 28° to the longitude of a station dedicated exclusively to SDO in New Mexico. This makes SDO with its HMI instrument one of the producers of solar magnetic field observation data that summarizes almost the entire 24th solar cycle.

Using the assumption that the data produced by HMI continues to increase, this study then chooses Random Ferns (RFe) as an algorithm to study and predict the occurrence of SFs. RFe is used to create an algorithm that still refers to the ensemble nature of Random Forest (RF) (Breiman, 2001), but is simpler. The implementation of Naive Bayesian Classification in RFe itself is known to optimally handle many features which is the key to increasing the level of classification. In this study, the implementation of RFe will be used to predict SFs which are divided into two scenarios, namely multiclass SFs and binary SFs. Both scenarios are adaptations of previous studies and at the same time a form of tuning in the use of RFe. In addition, this study also uses an oversampling method on training data to handle the imbalanced trait in SFs data.

2. PENELITIAN TERKAIT

Efforts to predict the emergence of SFs are broadly divided into two approaches, namely statistics (Barnes *et al*., 2007; Contarino *et al.,* 2009; Ternullo *et al*., 2006) and Machine Learning (Boucheron *et al*., 2015; Li *et al*., 2008; Nishizuka et al., 2017; Qahwaji and Colak, 2007; Yu et al., 2009). Reviewing the prediction efforts with a statistical approach, in his

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

research Barnes *et al* (2007) using ground-based vector magnetic field data obtained from the University of Hawai'i Imaging Vector Magnetograph. Barnes *et al* (2007) then applied a statistical discriminant analysis approach to predict the occurrence of SFs. Although the data used only covered a portion of a single solar cycle, the results showed performance comparable to the Bayesian approach and to the methods used by the Space Environment Center (SEC) of the United States.

Using solar magnetic vector field data sourced from SDO/HMI for four years and the ML Support Vector Machine (SVM) algorithm, Borba and Couvidat (2015) selected 25 parameters in 2071 Active Regions (ARs) and continued with the creation of a model to predict the occurrence of M and X classes in SFs (with M ≥ M1.0). Unlike other similar studies, which generally use ground-based data, this study is the first time that vector magnetogram data sourced from instruments in space and in large quantities is used to predict the occurrence of SFs. Borba and Couvidat (2015) then evaluated the model that had been created with an emphasis on TSS metrics and obtained relatively good results.

In contrast to Borba and Couvidat (2015), Liu *et al* (2017) conducted research to predict the occurrence of SFs using the ML Random Forest (RF) algorithm. Using the same data source as Borba and Couvidat (2015), Liu *et al (2017)* research used solar magnetic vector data in the range of May 2010-December 2016. The data is then selected based on the time of the last appearance of each day. After that, the undersampling method was used to handle the imbalanced nature of SFs data. Another thing that distinguishes the research of Liu *et al* (2017) from the research of Borba and Couvidat (2015) is the addition of scenarios from the predicted class, namely with the threshold ≥ B1.0 which has implications for the addition of multiclass scenarios (B, C, M, and X). Overall, there are two SFs prediction scenarios in Liu *et al* (2017) research , namely multiclass SFs (B, C, M, and X) and binary SFs (BC and MX). The conclusion of the research of Liu *et al* (2017) is that the use of HMI and RF parameters to predict SFs is a valid method and is able to provide good results.

3. METHODS

The data in this study consisted of flux data of SFs detected in the environment around the earth which was mapped with magnetic vector data on the appearance of SFs in the sun. SFs flux data in the environment around the earth is continuously detected by X-ray Sensors (XRS) found on the Geostationary Operational Environmental Satellite (GOES). Meanwhile, magnetic vector data is a data product from HMI instruments in SDO. Mapping flux SFs with magnetic vector data was carried out using the National Oceanic and Atmospheric Administration (NOAA) AR Numbers-HMI Active Region Patches (HARPs) Numbers mapping dictionary issued by the Joint Science Operations Center (JSOC) of Stanford University.

This study uses recommendations Borba and Couvidat (2015) about 13 parameters that have significance to the emergence of SFs. These 13 parameters are a combination of two data series contained in JSOC, namely hmi. SHARPs and cgem. Lorentz. The data is then transformed in stages as shown in **Table 1**. The details of the 13 parameters are found in **Table 2**. There are a total of 2256 SFs data used in this study with a composition of 30 class X, 443 class M, 1032 class C, and 751 class B for multiclass SFs; and 473 class MX and 1783 class BC for binary SFs.

Table 1. Stages of SFs data transformation.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

Transformation process	Description of the process
Labeling data SFs	In multiclass SFs scenarios, classes are simplified based on the letters of the class detected by XRS. For example, if an SF has class X5.0, it will be simplified to class X only. There are four classes in this scenario, namely B, C, M, and X. Meanwhile, in the SFs binary scenario, data with class B and class C is converted to class BC; and class M and class X were changed to class MX.
SFs data cleanup	Rows with one or more missing values from all 13 magnetic vector parameters are removed and not included in the next stage.
Daily SFs data selection	SFs data is selected based on the largest class that appears every day, except for class M and class X.

Table 2. The 13 parameters of the magnetic vector along with the source, description, and formula adapted from Borba and Couvidat (2015)

Magnetic Field Parameters	Source	Description	Formula
ABSNJZH	hmi.SHARPs	The absolute value of the net current helicity	$H_{c_{abs}} \propto \left\lvert \sum B_z \cdot J_z \right\rvert$
AREA_ACR	hmi.SHARPs	Areas of the strong pixel field in AR	$Area = \sum Pixels$
EPSZ	cgem.Lorentz	Number of Z-components of the normalized Lorentz force	$\delta F_z \propto \frac{\sum(B_x^2 + B_y^2 - B_z^2)}{\sum B^2}$
MEANPOT	hmi.SHARPs	Average photospherical magnet-free energy	$\bar{\rho} \propto \frac{1}{N}\sum(B^{Obs} - B^{Pot})^2$
R_VALUE	hmi.SHARPs	Amount of *flux* near the polarity inversion line	$\Phi = \sum \lvert B_{Los} \rvert dA \ (within\ R\ mask)$
SAVNCPP	hmi.SHARPs	Number of net current modulus per polarity	$J_{z_{sum}} \propto \left\lvert \sum^{B_z^+} J_z\, dA \right\rvert + \left\lvert \sum^{B_z^-} J_z\, dA \right\rvert$
SHRGT45	hmi.SHARPs	Area with an angle that shifts more than 45 degrees	$\frac{Area\ with\ Shear > 45^o}{total\ area}$
TOTBSQ	cgem.Lorentz	Total force of Lorentz force	$F \propto \sum B^2$
TOTFZ	cgem.Lorentz	Total Z-components of the Lorentz style	$F_z \propto \sum (B_x^2 + B_y^2 - B_z^2) dA$
TOTPOT	hmi.SHARPs	Total photosphere magnet-free energy density	$\rho_{tot} \propto \sum (\vec{B}^{Obs} - \vec{B}^{Pot})^2 dA$
TOTUSJH	hmi.SHARPs	Total unidentified helix currents	$H_{c_{total}} \propto \sum \lvert B_z \cdot J_z \rvert$
TOTUSJZ	hmi.SHARPs	Total unidentified vertical currents	$J_{z_{total}} = \sum \lvert J_z \rvert\, dA$
USFLUX	hmi.SHARPs	Total unidentified flux	$\Phi = \sum \lvert B_z \rvert dA$

After going through the data preprocessing stage, the next stage is the folding creation stage. At this stage, the data is formed into 5 folds, which means that 80% of the data will be training data, while 20% of the data will be test data.

Table 3. Experimental scenarios used in the study.

Scenario Name	Scenario
Number of ferns	200 ferns, 400 ferns, 600 ferns, 800 ferns, 1000 ferns, and 5000 ferns
Number of features	3 features, 5 features, 7 features, 9 features, 11 features, and 13 features

Table 3. (continue) Experimental scenarios used in the study.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

Scenario Name	Scenario
Sampling techniques	Oversample
Number of classes	Multiclass (B, C, M, and X) and binary (BC and MX)

The author does the rep at this stage 10 times, so there will be 10 sets of data for each scenario. A series of experiments were then carried out to obtain the most optimal SFs prediction models. Based on the characteristics of RFe, the author determines four factors as a form of tuning of the modeling process. Details of these four factors are found in After going through the data preprocessing stage, the next stage is the folding creation stage. At this stage, the data is formed into 5 folds, which means that 80% of the data will be training data, while 20% of the data will be test data.

Table 3 RFe-based SFs prediction model is then built using training data. Once the model is formed, the model is then evaluated using test data which is the remaining 20% chunk of the overall data.

Prediction of test data using the model that has been created will produce a list of class predictions which are then formed into a confusion matrix. After becoming a confusion matrix, the author then also measured the performance of the model that had been formed using several metrics, including sensitivity/recall, precision, and TSS. By using these measurement metrics, the results of this study can be compared with similar studies such as (Borba and Couvidat, 2015; Nishizuka et al., 2017; Liu et al., 2017; Yuan et al., 2010; Bloomfield et al., 2012; Ahmed et al., 2013). The details of the entire prediction process using RFe are as shown in **Figure 1**.

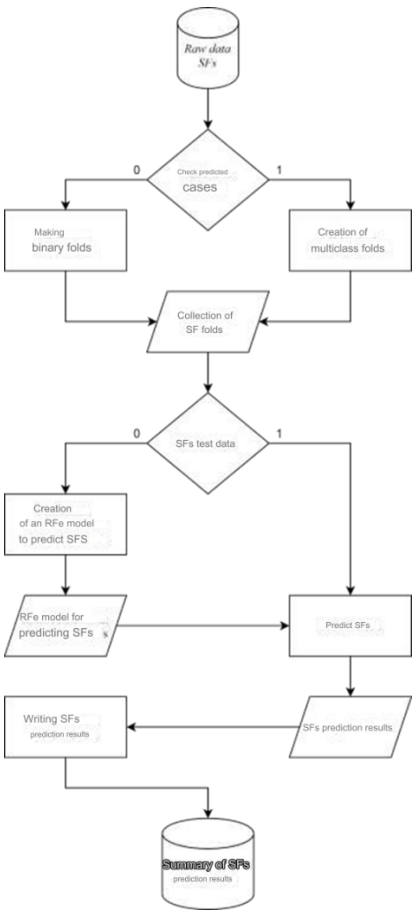

Figure 1. Rfe implementation flow for prefixing SFs.

3. RESULTS AND DISCUSSION

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

There were a total of 3575 experiments that included multiclass SFs and binary SFs scenarios. In the multiclass SFs scenario itself, the highest overall accuracy was recorded at 62.83% with details as shown in **Table 4**.

Table 4. The results of the SFs multiclass scenario experiment with the highest overall accuracy.

Description					Value	
Overall accuracy					0.6283	
Experiment number on multiclass 800 ferns					212	
Number of ferns					800	
Number of features					11	
Number of k-folds					5	
Number of fold					2	
Confusion Matrix						
Criteria	**B**	**C**	**M**	**X**	**Macro average**	**Weighted average**
Precision	0.619	0.740	0.552	0.182	0.523	0.656
Sensitivity	0.900	0.454	0.596	0.333	0.571	0.628
Specificity	0.725	0.865	0.882	0.980	0.863	0.824
F1-Score	0.734	0.563	0.573	0.235	0.526	0.617
TSS	0.625	0.319	0.477	0.313	0.434	0.452
Support	150	207	89	6	452	452

As for the SFs binary scenario, the highest overall accuracy was recorded at 85.17% with details as shown in **Table 5**.

Table 5. The results of the SFs binary scenario experiment with the highest overall.

Description				Value
Overall accuracy				0.8517
Experiment number on multiclass 800 ferns				253
Number of ferns				400
Number of features				13
Number of k-folds				5
Number of fold				3
Confusion Matrix				
Criteria	**BC**	**MX**	**Macro average**	**Weighted average**
Precision	0.948	0.609	0.778	0.876
Sensitivity	0.860	0.821	0.840	0.852
Specificity	0.821	0.860	0.840	0.829
F1-Score	0.902	0.700	0.801	0.859
TSS	0.681	0.681	0.681	0.681
Support	357	95	452	452

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

In addition, the plot results from the metrics in the SFs multiclass scenario as shown in **Figure 2** show that the increase in the number of features results in an upward trend in the percentage of the predicted results. This applies to almost every metric used, both macro averages and weighted averages. However, the trend of percentage increase from this prediction is not linear, because in the multiclass SFs scenario of 800 ferns, 1000 ferns, and 5000 ferns, the greatest overall accuracy does not fall on the largest number of features.

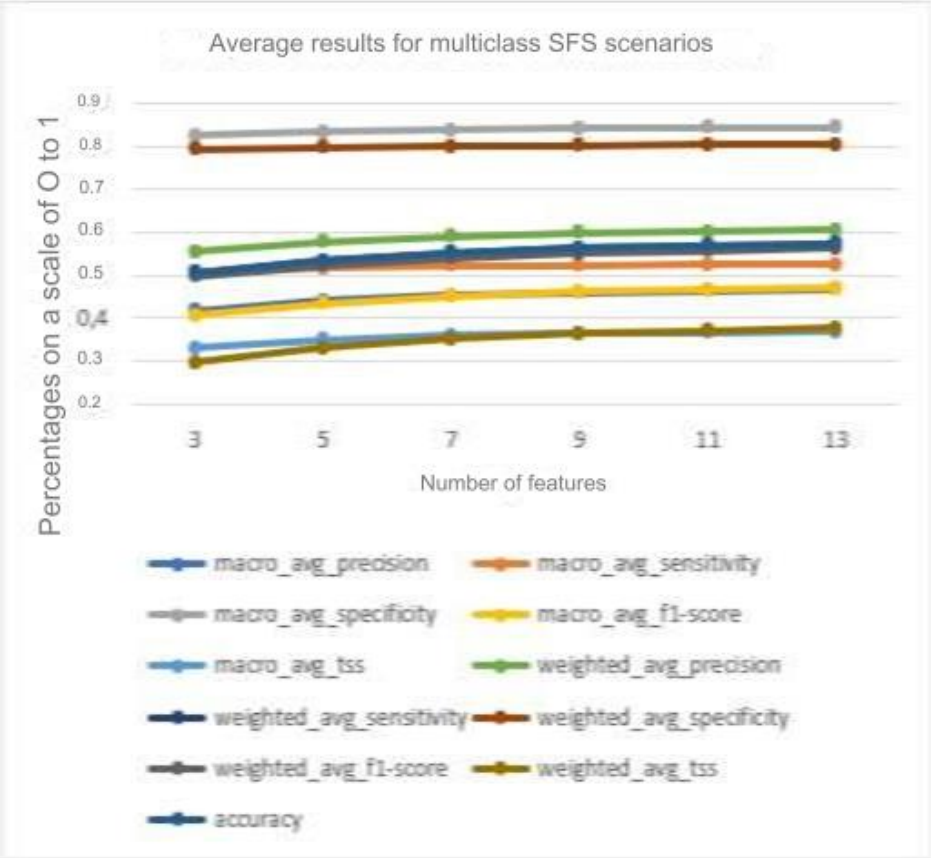

Figure 2. Line charts containing macro and weighted averages for precision, sensitivity, specificity, f1-score, TSS, and accuracy of experimental results for multiclass SFs scenarios.

This also applies to the SFs binary scenario where the results of the plot metrics show that the increase in the number of features results in an upward trend in the percentage of the prediction results as shown in **Figure 3**. The upward trend applies to almost every metric used, both macro averages and weighted averages. However, the trend of increasing the percentage of this prediction is also not linear, because in the binary SFs scenario of 200 ferns, 600 ferns, and 800 ferns, the greatest overall accuracy does not fall on the largest number of features.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

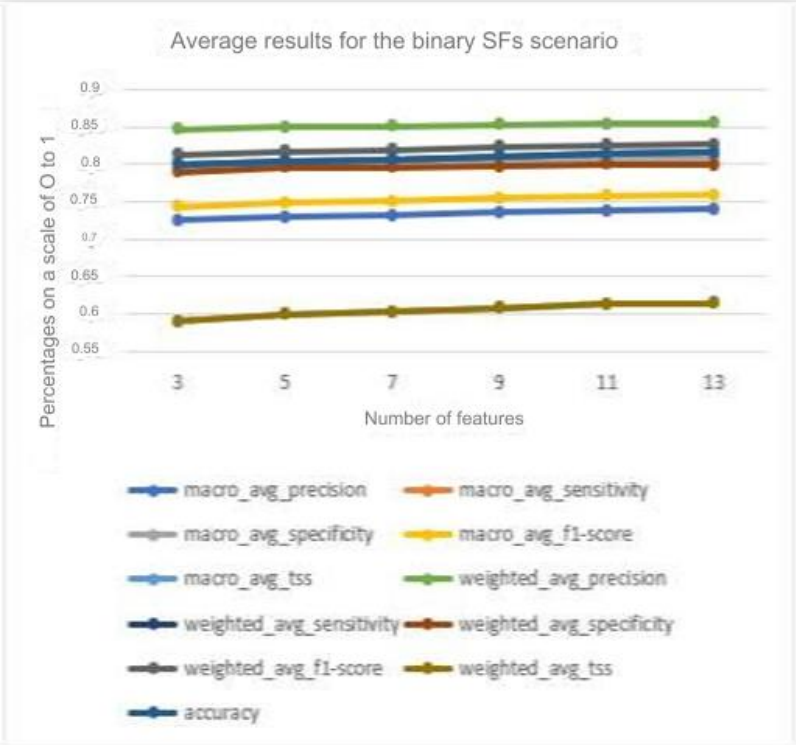

Figure 3. A line chart containing macro and weighted averages for precision, sensitivity, specificity, f1-score, TSS, and accuracy of experimental results for binary SFs scenarios.

4. CONCLUSION

Based on the results and discussion of the experiments that have been carried out, it can be concluded that the use of RFe to predict the occurrence of SFs using magnetic vector data is a valid method. Furthermore, the results of the study also show that the use of RFe with SFs data that covers almost the entire 24th solar cycle to predict the occurrence of SFs can outperform some aspects of measurement in previous studies. In addition, it can also be seen that the addition of the number of features or attributes of magnetic vector data into the modeling provides an upward trend in almost all classification performance measurement metrics in this study. Overall, although multiclass SFs B, C, M, X, predictions can be made, it can be seen that binary SFs BC and MX predictions are more optimal in classifying SFs based on magnetic vector data on the sun.

6. AUTHORS' NOTE

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

7. REFERENCES

Ahmed, O. W., Qahwaji, R., Colak, T., Higgins, P. A., Gallagher, P. T., and Bloomfield, D. S. (2013). Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Solar Physics, 283*, 157-175.

Barnes, G., Leka, K. D., Schumer, E. A., and Della-Rose, D. J. (2007). Probabilistic forecasting of solar flares from vector magnetogram data. *Space Weather, 5*(9).

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

Bloomfield, D. S., Higgins, P. A., McAteer, R. J., and Gallagher, P. T. (2012). Toward reliable benchmarking of solar flare forecasting methods. *The Astrophysical Journal Letters, 747*(2), L41.

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

Boucheron, L. E., Al-Ghraibah, A., and McAteer, R. J. (2015). Prediction of solar flare size and time-to-flare using support vector machine regression. *The Astrophysical Journal, 812*(1), 51.

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5-32.

Contarino, L., Zuccarello, F., Romano, P., Spadaro, D., Guglielmino, S. L., and Battiato, V. (2009). Flare forecasting based on sunspot-groups characteristics. *Acta Geophysica, 57*, 52-63.

Emslie, A. G., Dennis, B. R., Shih, A. Y., Chamberlin, P. C., Mewaldt, R. A., Moore, C. S., ... and Welsch, B. T. (2012). Global energetics of thirty-eight large solar eruptive events. *The Astrophysical Journal, 759*(1), 71.

Ghezzi, I., and Ruggles, C. (2007). Chankillo: a 2300-year-old solar observatory in coastal Peru. *Science, 315*(5816), 1239-1243.

Grätzel, M. (2007). Photovoltaic and photoelectrochemical conversion of solar energy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365*(1853), 993-1005.

Li, R., Cui, Y., He, H., and Wang, H. (2008). Application of support vector machine combined with K-nearest neighbors in solar flare and solar proton events forecasting. *Advances in Space Research, 42*(9), 1469-1474.

Liu, C., Deng, N., Wang, J. T., and Wang, H. (2017). Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm. *The Astrophysical Journal, 843*(2), 104.

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., and Ishii, M. (2017). Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. *The Astrophysical Journal, 835*(2), 156.

Priest, E. R., and Forbes, T. G. (2002). The magnetic nature of solar flares. *The Astronomy and Astrophysics Review, 10*(4), 313-377.

Qahwaji, R., and Colak, T. (2007). Automatic short-term solar flare prediction using machine learning and sunspot associations. *Solar Physics, 241*, 195-211.

Ruban, A. V. (2015). Evolution under the sun: Optimizing light harvesting in photosynthesis. *Journal of experimental botany, 66*(1), 7-23.

Ternullo, M., Contarino, L., Romano, P., and Zuccarello, F. (2006). A statistical analysis of sunspot groups hosting M and X flares. *Astronomische Nachrichten: Astronomical Notes, 327*(1), 36-43.

Yu, D., Huang, X., Hu, Q., Zhou, R., Wang, H., and Cui, Y. (2009). Short-term solar flare prediction using multiresolution predictors. *The Astrophysical Journal, 709*(1), 321.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

Yuan, Y., Shih, F. Y., Jing, J., and Wang, H. M. (2010). Automated flare forecasting using a statistical learning technique. *Research in Astronomy and Astrophysics, 10*(8), 785.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088