Interpretable Deep Learning for Solar Flare Prediction

Vinay Ram Gazula
Department of Data Science
New Jersey Institute of Technology
Newark, NJ 07102, USA
vg472@njit.edu

Katherine G. Herbert
School of Computing
Montclair State University
Montclair, NJ 07043, USA
herbertk@montclair.edu

Yasser Abduallah, Jason T. L. Wang
Department of Computer Science
New Jersey Institute of Technology
Newark, NJ 07102, USA
ya54@njit.edu, wangj@njit.edu

***Abstract*—We propose to incorporate three interpretable methods, namely SHAP (SHapley Additive exPlanations), PDP (partial dependence plots) and Anchors, into a deep learning-based model, called SolarFlareNet, for operational flare forecasting. SolarFlareNet takes as input a sample of SHARP (Space-weather HMI Active Region Patches) magnetic parameters and predicts as output whether a solar flare would occur within the next 24 hours. We analyze flare events that occurred from May 2010 to December 2022 using the Geostationary Operational Environmental Satellite's X-ray flare catalogs and construct a database of flares with identified active regions in the catalogs. This database, together with the SHARP magnetic parameters, is used to train and test the SolarFlareNet model. Our experimental results describe the use of the three proposed methods (SHAP, PDP, and Anchors) to interpret the SolarFlareNet model and demonstrate the effectiveness of the methods.**

***Index Terms*—Solar flares, Deep learning, Interpretability**

I. Introduction

Space weather (SWx) refers to transients in the space environment traveling from the Sun through the heliosphere to Earth. In the recent decade, tackling the difficult task of understanding and forecasting violent solar eruptions including flares and coronal mass ejections, which are sources of SWx, and their terrestrial impacts, has become a strategic national priority, as SWx affects the life of human beings, including communication, transportation, power supplies, national defense, space travel, and more. A flare appears as a sudden, intense brightening of an active region (AR) on the Sun and will remain for a duration ranging from several minutes to hours. Flares emit a wide range of electromagnetic radiation, including X-rays and ultraviolet (UV) light, and also release high-energy particles such as protons, electrons and alpha particles, which can enter the Earth's atmosphere along the open magnetic field lines near the polar caps.

In general, flares are categorized into A, B, C, M, and X classes, from weakest to strongest, based on their peak soft X-ray flux in the $1-8$ angstrom ($0.1-0.8$ nm) channel measured by the Geostationary Operational Environmental Satellite. When solar activity is strong and an intense solar flare occurs pointed towards Earth, it can possibly cause a radio blackout, a solar radiation storm, or a geomagnetic storm, all of which have adverse effects on the infrastructure in space and technological systems on Earth. Being able to forecast solar flares helps reduce their damage to Earth.

The triggering mechanism of a flare is not yet fully understood. Past studies indicate that the intensity of flares can be determined by several photospheric vector magnetogram features, including the size, structure, topology, and complexity of active regions (ARs), unsigned magnetic flux, gradient of the magnetic field, magnetic energy dissipation, vertical electric currents, integrated Lorentz forces, magnetic shear, magnetic helicity injection, etc. [1], [2]. Although substantial effort has been spent, the physical relationships between flare productivity and ARs are far from understood. With the increase in flare-related data [3] and advances in machine learning methods, researchers started implementing different machine learning models to predict flares and associated eruptive events such as coronal mass ejections and solar energetic particles.

Recently, Abduallah et al. [4] used the SHARP parameters to build an operational, near-real-time flare forecasting system, named SolarFlareNet, implemented with a deep learning model. The SolarFlareNet system can predict whether there would be a γ-class flare within 24 to 72 hours, where γ is $\geq$M5.0, $\geq$M or $\geq$C. In this paper, we extend the SolarFlareNet model by adding interpretability to the model using three methods, namely SHAP (SHapley Additive exPlanations), PDP (partial dependence plots) and Anchors, to understand feature significance and identify the features that influence the model's predictions most. Here, we focus on the prediction of $\geq$M-class flares that would occur within the next 24 hours, where a $\geq$ M-class flare refers to an M- or X-class flare.

A. Related Work

Interpretable machine learning, or explainable artificial intelligence (XAI), has been incorporated into solar flare predictions. For example, Feldhaus and Carande [5] implemented LIME (Local Interpretable Model-Agnostic Explanations), a widely used XAI tool, into a solar flare prediction model built using support vector machines (SVM). Li et al. [6] developed a model named FAST-CF, which provides intuitive, post-hoc counterfactual explanations for solar flare predictions. Pandey et al. [7] performed a post hoc analysis on a deep learning-based full-disk solar flare prediction model using three methods, including class activation mapping, deep SHAP,

2375-0197/24/$31.00 ©2024 IEEE
DOI 10.1109/ICTAI62512.2024.00078

and integrated gradients. In contrast to the above studies, our work focuses on incorporating SHAP, PDP, and Anchors into a transformer-based model for operational flare forecasting using SHARP magnetic parameters. Other related methods can be found in [8], [9].

The remainder of this paper is organized as follows. Section II describes the data used in our study. Section III details the three interpretable methods (SHAP, PDP, and Anchors) implemented in the SolarFlareNet model described in [4]. Section IV reports the experimental results. Section V concludes the paper.

II. Data

We survey the flare events that occurred from May 2010 to December 2022 using the Geostationary Operational Environmental Satellite's X-ray flare catalogs provided by the National Centers for Environmental Information (NCEI) to build a flare database with identified active regions (ARs) in the NCEI flare catalogs. This database is used to construct labels of the data samples suitable for machine learning. SolarFlareNet utilizes SHARP magnetic parameters [10] acquired from the Joint Science Operations Center (JSOC), which can be accessed at http://jsoc.stanford.edu/. On the JSOC website, the data samples, which are composed of SHARP parameters at a cadence of 12 minutes from the `hmi.sharp_cea_720s` series, are downloaded with the Python package "SunPy" [11]. These data samples are used to forecast solar flares.

Following previous studies [1], [12]–[15], nine magnetic parameters are considered in the SHARP data series. These nine parameters include the total unsigned current helicity (TOTUSJH), total unsigned vertical current (TOTUSJZ), total unsigned flux (USFLUX), mean characteristic twist parameter (MEANALP), sum of flux near polarity inversion line (R_VALUE), total photospheric magnetic free energy density (TOTPOT), sum of the modulus of the net current per polarity (SAVNCPP), area of strong field pixels in the active region (AREA_ACR), and absolute value of the net current helicity (ABSNJZH). The details of the formulas used to calculate the features can be found in previous work [1], [12]–[14].

These SHARP magnetic parameters are derived from photospheric magnetic field data taken by the Helioseismic and Magnetic Imager on board the Solar Dynamics Observatory (SDO). The parameters have been available since 2010 to the present day. The parameters and related data products have been studied by many researchers. They are useful in such areas as quantifying electric currents and flux cancelation within solar ARs, connecting photospheric magnetic field properties to the kinematic properties of coronal mass ejections, developing high-resolution magnetohydrodynamic simulations of ARs as they evolve over time, and solar eruption predictions, to name a few [16].

III. Methodology

Understanding and interpreting the predictions made by the SolarFlareNet model is challenging due to its complex model architecture. However, understanding the predictions made by the model is important, as a potential flare outcome could be devastating. In an effort to make the SolarFlareNet model interpretable and reliable, three methods are integrated into the model, including deep SHAP, PDP, and Anchors.

Deep SHAP. SHAP (SHapley Additive exPlanations) [17] is a widely used method for explaining individual predictions made by machine learning models. The method provides a clear and intuitive understanding of why a model made a particular prediction by attributing the prediction to its input features. The contribution of each feature is calculated, often represented by SHAP values, using cooperative game theory. The method can also be used to explain the predictions for all test data samples in a global way. Deep SHAP [17], a combination of the DeepLIFT algorithm [18] and the SHAP values, can be used to extract additional information from deep learning models. The SHAP Python package includes a "DeepExplainer" class, an enhanced version of the DeepLIFT algorithm, which is used to approximate the conditional expectations of SHAP values using a selection of background samples for deep learning models.

PDP. Partial dependence plots (PDP) help visualize the marginal effects one or two features have on the predicted outcome of a model. Using PDP, we can determine whether the relationship between the target and feature is linear, monotonic or more complex. If the features are not correlated to each other, then PDP can reflect how the features influence the model prediction on average [19]. The key advantage of using PDP is that the computation is quite intuitive, meaning that even a layman can easily understand the idea behind PDP.

Anchors. Anchors [20] is a model-agnostic method that interprets the behavior of a complex model using high-precision rules, known as anchors, representing local, "sufficient conditions" for a prediction. When changes in other feature values have no effect on the prediction, a rule anchors it. The Anchors method uses reinforcement learning techniques in conjunction with a graph search algorithm to minimize the number of model calls while recovering from local optima [19]. The method generates a set of decision rules for a given instance that anchors the prediction of a black-box model. These anchors may vary depending on the given instance; for example, if the given instance is close to the decision boundary of the model prediction function, then the anchors may be complex and provide less coverage.

IV. Experiments and Results

We used 170,193 training data samples to retrain the SolarFlareNet model after the interpretable methods including Deep SHAP, PDP, and Anchors are incorporated into the model. We then used 47,775 test data samples to test the model and calculate performance metrics. The accuracy of the newly trained model is 0.907, which is close to the accuracy of the original model described in [4], indicating that there is no significant loss in model accuracy. The interpretable tools are then tested using a smaller, randomly selected subset of the test set, for which the results are presented in this section.

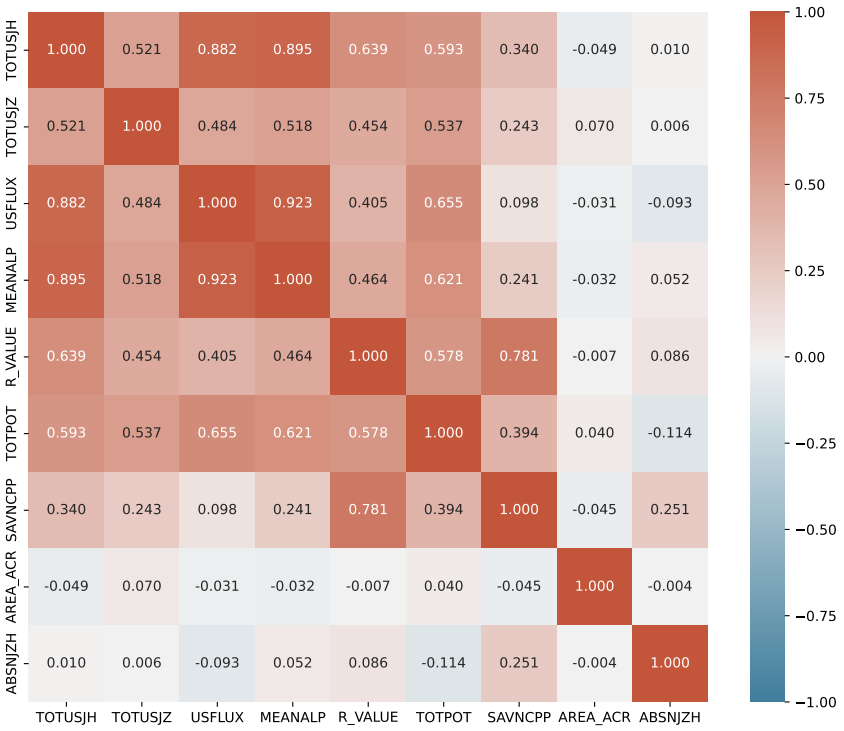


Fig. 1. The Pearson correlation coefficient matrix of the nine features used in our study.

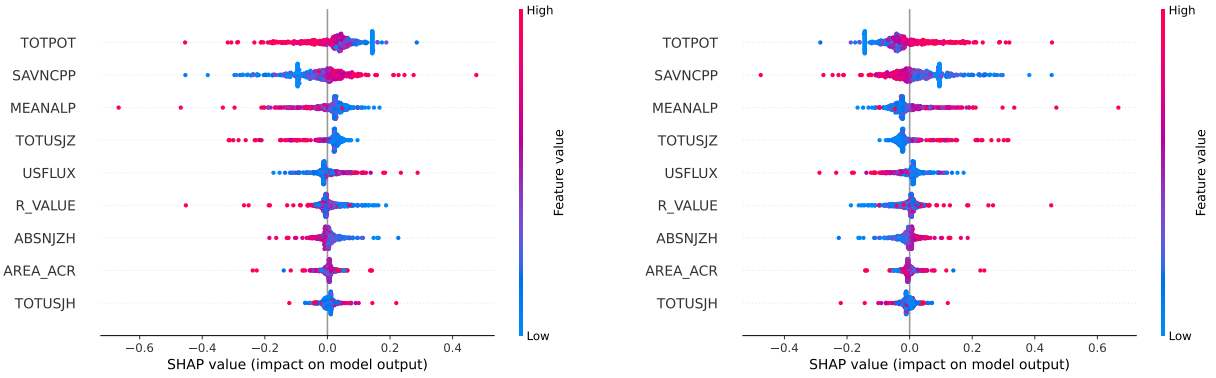
To understand the interactions of the nine features described in Section II, we first calculate the Pearson correlation coefficients between the features. Figure 1 shows the Pearson correlation coefficient matrix of the features. Red and blue represent positive and negative correlation coefficients approaching 1 and -1, respectively. The darker the color (either red or blue), the closer the absolute value of the correlation coefficient is to 1. A value of 1 means that a linear equation perfectly describes the relationship between two features X and Y, where all feature values are on a line for which Y increases as X increases. A value of -1 means that all feature values lie on a line for which Y decreases as X increases. A zero value means that there is no linear correlation between the features X and Y.

For example, consider the TOTPOT feature. Figure 1 shows that USFLUX is most correlated with TOTPOT with coefficient value 0.655 while AREA_ACR is least correlated with TOTPOT with coefficient value 0.040. Consider the TOTUSJH feature. Figure 1 shows that MEANALP is most correlated with TOTUSJH with coefficient value 0.895 while ABSNJZH is least correlated with TOTUSJH with coefficient value 0.010.

A. Model Interpretation Using Deep SHAP

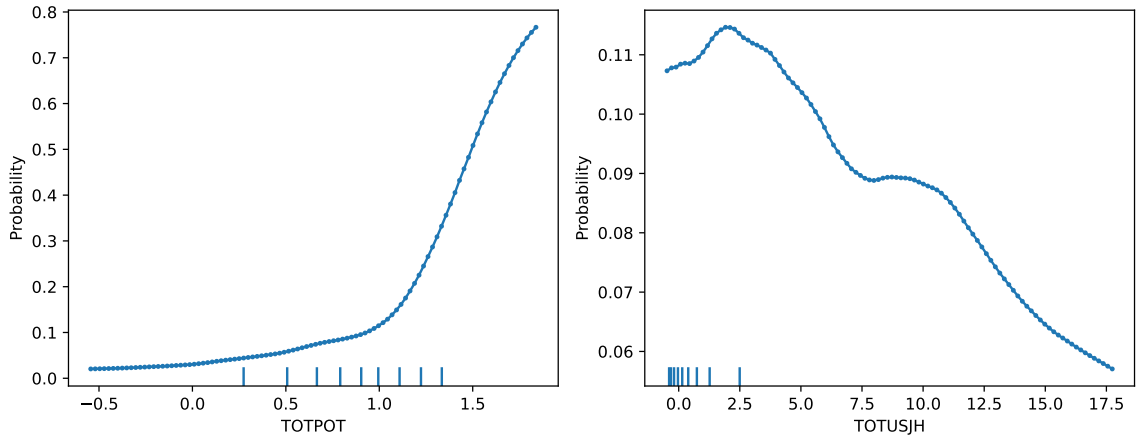
Figure 2 displays the SHAP summary plots for all data samples in the test set. Class 0 means that the model predicts that there will be no flare event in 24 hours. Class 1 means that the model predicts that there will be a flare event during the next 24-hour period. The plot in Figure 2(a) is for class 0 and the plot in Figure 2(b) is for class 1. SHAP values are displayed on the x-axis. Each point in a summary plot represents the SHAP value of a test data sample with respect to a feature. The negative SHAP value means a negative effect. The positive SHAP value means a positive effect. The overlap points give us an idea of how the SHAP values are distributed per feature. The color denotes the value of a feature (that is, the value of a magnetic parameter), ranging from low to high.

Figure 2(a) and Figure 2(b) are opposite. For example, consider the TOTPOT feature. The high (red) values of this feature produce negative SHAP values with negative effects for predicting class 0 (that is, there will be no flare within 24 hours), as shown in Figure 2(a). However, these high (red) values of the feature produce positive SHAP values with positive impacts for predicting class 1 (that is, there will be a flare within 24 hours), as shown in Figure 2(b). On the other hand, the low (blue) values of this feature produce positive SHAP values with positive impacts for predicting class 0 as shown in Figure 2(a). These low (blue) values of the feature produce negative SHAP values with negative impacts for predicting class 1 as shown in Figure 2(b). Thus, when



(a) (b)

Fig. 2. SHAP summary plots for (a) class 0, and (b) class 1, respectively.



(a) (b)

Fig. 3. 1-D partial dependence plots for (a) TOTPOT, and (b) TOTUSJH, respectively.

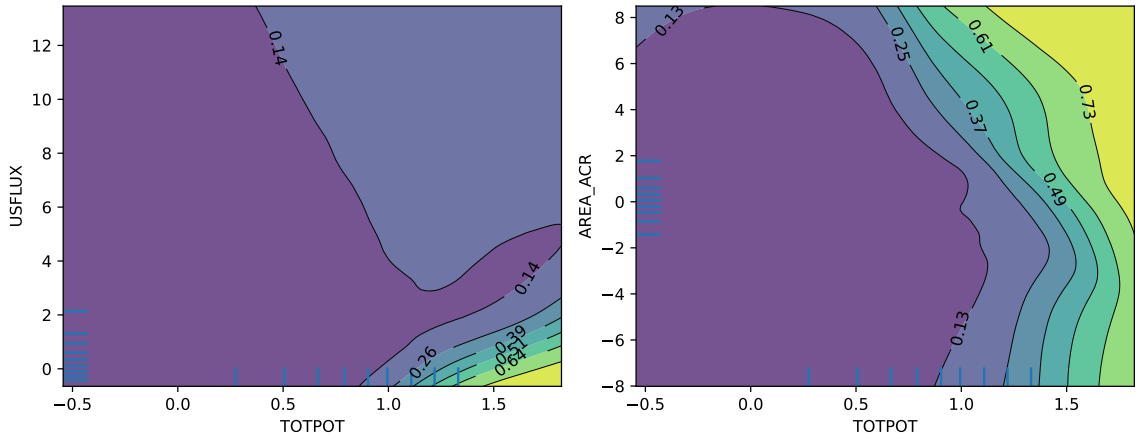
the model sees the high (red) values of the feature, the model tends to predict class 1; when the model sees the low (blue) values of the feature, the model tends to predict class 0.

The significance of a feature determines its location on the y-axis, with the most significant feature displayed at the top of the y-axis. According to Figure 2, TOTPOT is the most significant feature, while TOTUSJH is the least significant feature.

B. Model Interpretation Using PDP

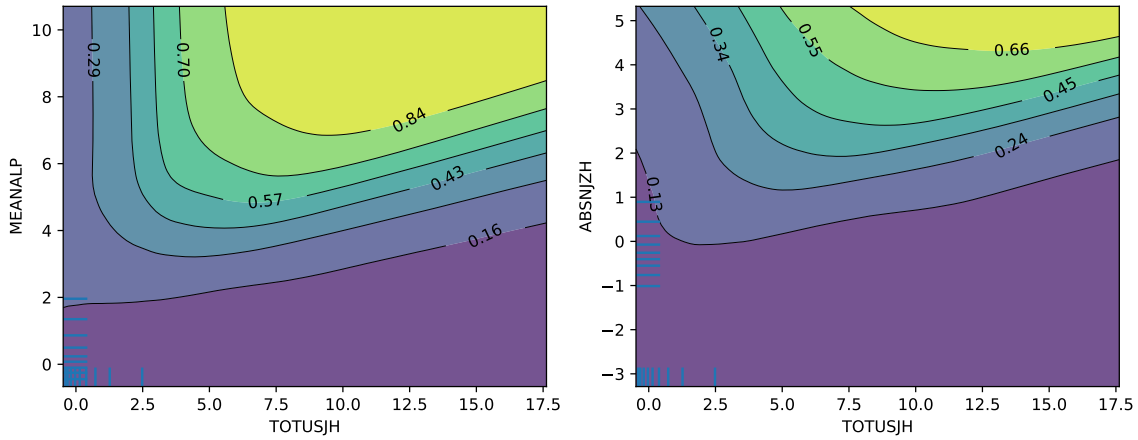
We further interpret the SolarFlareNet model by implementing PDP for the most significant feature (TOTPOT) and the least significant feature (TOTUSJH). Figure 3 displays 1-dimensional (1-D) partial dependence plots, showing the relationship between the value of the most/the least significant feature (along the x-axis) and the prediction probability (along the y-axis). The prediction probability represents the probability that a solar flare will occur within the next 24 hours. The distribution of the feature values is represented by the blue lines on the x-axis where each bin size represents 10% of all feature values. Figure 3(a) corresponds to the TOTPOT feature, which is considered the most significant feature. By analyzing the plot, we can see that the TOTPOT feature values are spread uniformly, and the prediction probability rapidly increases as the feature value increases over the value 1.0. Thus, larger TOTPOT feature values imply a higher probability that a flare will occur within 24 hours. Figure 3(b) corresponds to the TOTUSJH feature, which is considered the least significant feature. We see that 90% of its feature values are less than 2.5. Furthermore, these small TOTUSJH feature values imply a low probability for a flare to occur within 24 hours.

We can also visualize the partial dependence of two features at once to understand how these features interact with each other and affect the model prediction. For example, consider again the most significant TOTPOT feature and the least significant TOTUSJH feature. Figure 4 shows the 2-D partial dependence plots between TOTPOT and its most



(a) (b)

Fig. 4. 2-D partial dependence plots for TOTPOT and its (a) most correlated feature USFLUX, and (b) least correlated feature AREA_ACR, respectively.



(a) (b)

Fig. 5. 2-D partial dependence plots for TOTUSJH and its (a) most correlated feature MEANALP, and (b) least correlated feature ABSNJZH, respectively.

correlated feature USFLUX and its least correlated feature AREA_ACR, respectively. Figure 5 shows the 2-D partial dependence plots between TOTUSJH and its most correlated feature MEANALP and its least correlated feature ABSNJZH, respectively. The contours in the 2-D partial dependence plots represent the model's prediction probabilities for the corresponding feature values. The distribution of feature values is represented by the blue lines on the x-axis and y-axis for the respective/corresponding features.

In Figure 4(a), for TOTPOT values $\geq$ 1.0 and USEFLUX values ranging from -1 to 2, the model's prediction probabilities vary a lot, with high TOTPOT values and low USEFLUX values resulting in high prediction probabilities (represented by yellow color). Thus, when the model sees high TOTPOT values and low USEFLUX values, the model tends to predict that there will be a flare within 24 hours. In Figure 4(b), the prediction probability is not affected by changes in the AREA_ACR values when the TOTPOT values are less than 1.0. The prediction probability increases only when the TOTPOT values are $\geq$ 1.0.

In Figure 5, we see that the prediction probability is not affected at all by changes in TOTUSJH values, attesting that TOTUSJH is the least significant feature. The model prediction here is only affected by the changes in the values of MEANALP and ABSNJZH, respectively. When the model sees high values of MEANALP and ABSNJZH, the model tends to predict that there will be a flare within 24 hours.

C. Model Interpretation Using Anchors

In contrast to Deep SHAP and PDP described above, which are used to explain the predictions for all test data samples,

IF $-0.55 <$ ABSNJZH ≤ -0.40 **AND**
$1.27 <$ TOTUSJH ≤ 2.50 **AND**
$0.91 <$ SAVNCPP ≤ 1.04 **AND**
$1.35 <$ MEANALP ≤ 1.96 **AND**
$1.02 <$ AREA_ACR ≤ 1.78 **AND**
$1.15 <$ TOTUSJZ ≤ 2.37 **AND**
$1.09 <$ R_VALUE ≤ 1.54 **AND**
TOTPOT > 1.33 **AND**
$1.31 <$ USFLUX ≤ 2.13
THEN PREDICT Flare $= 0$

Fig. 6. Anchors' explanation for a negative flare prediction.

IF $-0.26 <$ ABSNJZH ≤ -0.07 **AND**
TOTUSJH > 2.50 **AND**
$0.77 <$ SAVNCPP ≤ 0.91 **AND**
MEANALP > 1.96 **AND**
AREA_ACR ≤ -1.42 **AND**
TOTUSJZ > 2.37 **AND**
$1.54 <$ R_VALUE ≤ 2.29 **AND**
TOTPOT > 1.33 **AND**
USFLUX > 2.13
THEN PREDICT Flare $= 1$

Fig. 7. Anchors' explanation for a positive flare prediction.

the Anchors method is used to explain individual prediction instances. Here, two individual instances, including a negative solar flare prediction and a positive solar flare prediction, are used to generate anchors. Since all nine features used in training the SolarFlareNet model are continuous numerical variables, the decision rules generated by Anchors are represented as ranges of feature values. Figure 6 displays the anchors for a negative flare prediction instance. Figure 7 displays the anchors for a positive flare prediction instance. These anchors highlight what feature values lead to a positive (negative, respectively) flare prediction instance. The rules help us better understand the cause of a decision.

V. Conclusion

In this paper, we incorporate XAI tools including Deep SHAP, PDP and Anchors into an operational flare forecasting system (SolarFlareNet) to predict whether a $\geq$M class flare would occur within the next 24 hours. In Deep SHAP, summary plots are used to rank features according to their significance or influence on the model predictions and how the feature values relate to the model predictions, respectively. PD plots (PDP) are used to determine the effects of single features and feature pairs on the predictions made by the SolarFlareNet system. Anchors are used to generate a set of decision rules in which the model predictions are anchored to either a negative or a positive flare prediction outcome. Together, the three tools explain the internal workings of the SolarFlareNet system while maintaining its accuracy.

References

[1] H. Liu, C. Liu, J. T. L. Wang, and H. Wang, "Predicting solar flares using a long short-term memory network," *The Astrophysical Journal*, vol. 877, no. 2, p. 121, 2019.

[2] Y. Abduallah, J. T. L. Wang, Y. Nie, C. Liu, and H. Wang, "DeepSun: Machine-learning-as-a-service for solar flare prediction," *Research in Astronomy and Astrophysics*, vol. 21, no. 7, p. 160, Aug. 2021.

[3] M. K. Georgoulis, D. S. Bloomfield, M. Piana, A. M. Massone, M. Soldati, P. T. Gallagher, E. Pariat, N. Vilmer, E. Buchlin, F. Baudin *et al.*, "The flare likelihood and region eruption forecasting (FLARECAST) project: Flare forecasting in the big data & machine learning era," *Journal of Space Weather and Space Climate*, vol. 11, p. 39, 2021.

[4] Y. Abduallah, J. T. L. Wang, H. Wang, and Y. Xu, "Operational prediction of solar flares using a transformer-based framework," *Scientific Reports*, vol. 13, no. 1, p. 13665, 2023.

[5] C. Feldhaus and W. Carande, "Explainable artificial intelligence for solar flare prediction," in *AGU Fall Meeting Abstracts*, vol. 2021, 2021, pp. NG45B–0576.

[6] P. Li, O. Bahri, S. F. Boubrahimi, and S. M. Hamdi, "Fast counterfactual explanation for solar flare prediction," in *Proceedings of the 21st IEEE International Conference on Machine Learning and Applications*, 2022, pp. 1238–1243.

[7] C. Pandey, R. A. Angryk, M. K. Georgoulis, and B. Aydin, "Explainable deep learning-based solar flare prediction with post hoc attention for operational forecasting," in *Proceedings of the 2023 International Conference on Discovery Science*, 2023, pp. 567–581.

[8] G. V. Datla, H. Jiang, and J. T. L. Wang, "An interpretable LSTM network for solar flare prediction," in *Proceedings of the 35th IEEE International Conference on Tools with Artificial Intelligence*, 2023, pp. 526–531.

[9] Y. Abduallah, V. R. Gazula, and J. T. L. Wang, "An interpretable transformer model for operational flare forecasting," in *Proceedings of the 37th International Florida Artificial Intelligence Research Society Conference*, 2024.

[10] M. G. Bobra, X. Sun, J. T. Hoeksema, M. Turmon, Y. Liu, K. Hayashi, G. Barnes, and K. Leka, "The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs–space-weather HMI active region patches," *Solar Physics*, vol. 289, pp. 3549–3578, 2014.

[11] S. J. Mumford, S. Christe, D. Pérez-Suárez, J. Ireland, A. Y. Shih, A. R. Inglis, S. Liedtke, R. J. Hewett, F. Mayer, K. Hughitt *et al.*, "SunPy—Python for solar physics," *Computational Science & Discovery*, vol. 8, no. 1, p. 014009, 2015.

[12] M. G. Bobra and S. Couvidat, "Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm," *The Astrophysical Journal*, vol. 798, no. 2, p. 135, 2015.

[13] M. G. Bobra and S. Ilonidis, "Predicting coronal mass ejections using machine learning methods," *The Astrophysical Journal*, vol. 821, no. 2, p. 127, 2016.

[14] C. Liu, N. Deng, J. T. L. Wang, and H. Wang, "Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm," *The Astrophysical Journal*, vol. 843, no. 2, p. 104, 2017.

[15] H. Liu, C. Liu, J. T. L. Wang, and H. Wang, "Predicting coronal mass ejections using SDO/HMI vector magnetic data products and recurrent neural networks," *The Astrophysical Journal*, vol. 890, no. 1, p. 12, 2020.

[16] M. G. Bobra, P. J. Wright, X. Sun, and M. J. Turmon, "SMARPs and SHARPs: Two solar cycles of active region data," *The Astrophysical Journal Supplement Series*, vol. 256, no. 2, p. 26, 2021.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 3145–3153.

[19] C. Molnar, *Interpretable Machine Learning*. lulu.com, 2020.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.