Disentangling Data Availability and Class Variability in Multivariate Time Series for Rare Event Prediction: A GAN-Based Approach to Solar Flare Forecasting

Junzhi Wen * Rafal A. Angryk †

Abstract

Accurate prediction of rare events is a longstanding challenge in many domains, primarily due to (1) severe impact of class imbalance and (2) the limited availability of critical minority class samples. In this work, we investigate the impact of data availability versus class variability on rare event prediction using solar flare forecasting as a case study. Leveraging the SWAN-SF dataset, we explore three training strategies within a generative adversarial network (GAN) framework. Our approach isolates the influence of class distribution by training GAN models on data from individual classes, and the discriminator is subsequently employed as a binary classifier to differentiate between strong flare events (MX) and non-flaring or low-level activity (NBC), which reflects the most need of operational community for solar flare forecasting. Experimental results reveal that neither the quantity nor the inherent variability of training samples significantly affect model performance, challenging our assumption that extreme events are more diverse and thus harder to predict. Our findings suggest that the difference in class variability between MX and NBC is not as significant as expected, and that data availability does not necessarily outweigh data variability in machine learning-based solar flare prediction.

1 Introduction.

Rare event prediction is inherently challenging due to severe class imbalance, where the events of interest occur far less frequently than common events. This imbalance complicates model training and evaluation, as the scarcity of minority class samples can lead to biased models that underperform on the rare, yet critical, events [6]. Traditionally, researchers have addressed this issue through oversampling of the minority class or undersampling of the majority class or more advanced mixtures of both approaches [4, 1, 29, 28, 16]. More recently, generative adversarial networks (GANs) [15] have emerged as a promising approach for data augmentation in imbalanced settings, primarily by generating synthetic samples for the minority class [2, 7, 8, 9, 10, 13].

However, the use of GANs for rare event prediction faces two major challenges. First, the limited availability of minority class data constrains the GAN training process, potentially hindering its ability to generate high-quality synthetic samples. Second, it is easy for people to assume that extreme natural events exhibit greater variance due to their rarity, which makes them more challenging to predict as their diverse characteristics may not be fully captured by synthetic data generation methods, especially when data is scarce. These challenges raise a question: can we leverage the abundance and presumed lower variability of common (majority) data to improve rare event prediction?

In this study, we investigate two aspects of rare event prediction. First, we explore the difference in class variability between rare and common event data and assess whether this difference, when data availability is the same, leads to divergent predictive performance. Second, we analyze the impact of common event data availability and determine whether it plays a more critical role in predictive performance for rare events. To address these questions, we employ a GAN framework to disentangle the effects of data availability and class variability by training models separately on data from a single class. The discriminator from the trained GAN is then directly used as a classifier to evaluate performance on testing data.

Solar flare prediction represents a compelling case study for rare event prediction due to the significant societal impacts of major solar flare events and the inherent challenges posed by their rarity. In our work, we utilize the SWAN-SF dataset [3], which consists of multivariate time series data of solar activities from solar cycle 24, to investigate these issues in the context of extreme natural events. By exploring various training strategies on SWAN-SF, we aim to shed light on the interplay between data availability and class variability in enhancing the performance of rare event prediction models.

*Georgia State University, Atlanta, GA, USA. Email: jwen6@student.gsu.edu

†Georgia State University, Atlanta, GA, USA. Email: angrykr@gmail.com

The remainder of this paper is organized as follows. Section 2 presents an overview of related work on solar flare prediction and the use of GANs to support rare event prediction. Section 3 describes the SWAN-SF dataset, our GAN architecture and training pipeline, and the evaluation metrics used in our experiments. In Section 4, we detail the experimental setups and discuss the results. Finally, Section 5 concludes the paper and suggests directions for future work.

2 Related Work.

Solar flare prediction has attracted significant attention over the past decades, with a wide variety of methods proposed to forecast flare occurrences. Early approaches relied on statistical models and physical simulations [30, 31, 14], while more recent work has increasingly adopted machine learning techniques [26, 5, 20, 19, 23]. These methods range from traditional algorithms, such as k-nearest neighbors (KNN) [26], support vector machines (SVM) [5], and random forests (RF) [20], to advanced neural networks, including convolutional neural networks (CNN) [23] and long short-term memory (LSTM) networks [19] that capture temporal dependencies in space weather data. However, a persistent challenge in this domain is the severe class imbalance, as major flare events are extremely rare compared to non-flaring or low-level activity, complicating both model training and evaluation.

To address the imbalance in solar flare and other rare-event datasets, several traditional techniques have been explored, including oversampling the minority class, undersampling the majority class, and employing synthetic data generation techniques such as SMOTE [6] and deep learning models [7, 8, 12] . More recently, generative adversarial networks (GANs) have emerged as a promising tool to deal with the class imbalance issue in rare event prediction. For example, [25] introduced an unsupervised approach to anomaly detection in medical imaging by training a GAN to model the distribution of healthy data, where anomalies are identified based on high reconstruction errors that indicate deviation from the learned manifold. Similarly, [18] leverages both the generator and discriminator of a trained GAN to detect anomalous multivariate time series data from real-world GPS datasets, using an anomaly score derived from discrimination and reconstruction. Furthermore, [8] addresses the class imbalance in solar flare prediction by generating synthetic samples of the minority class using a conditional GAN (CGAN) [21], thereby balancing the samples from the minority and majority classes during training.

While GAN-based approaches have demonstrated promise in refining decision boundaries and mitigating the effects of class imbalance for rare event prediction, to our knowledge no study has systematically compared the impact of data availability versus data variability when addressing class imbalance. In this work, we investigate how the distribution and availability of training samples affect the discriminator's ability to classify rare events, exploring different training strategies as detailed in Section 4.1.

Table 1: Class distribution and imbalance ratio of SWAN-SF across different partitions.

Partition	Class					Imbalance Ratio (MX : NBC)
	X	M	C	B	N	
1	165	1,089	6,416	5,692	60,130	1:58
2	72	1,392	8,810	4,978	73,368	1:62
3	136	1,288	5,639	685	34,762	1:29
4	153	1,012	5,956	846	43,294	1:43
5	19	971	5,753	5,924	62,688	1:75

3 Methodology.

3.1 SWAN-SF Dataset. The experiments in this study are based on the SWAN-SF dataset [3], a multivariate time series (MVTS) dataset developed for space weather data analytics, covering flare reports over an eight-year period (May 2010 – December 2018) during solar cycle 24. The dataset is constructed using a sliding-window methodology, where each MVTS sample is generated with a 12-hour observation window followed by a 24-hour prediction window. With a 12-minute cadence, each MVTS sample comprises 60 time steps of 51 predictive parameters, capturing the dynamic behavior of solar activities. Each sample is labeled with the strongest flare event occurring within the prediction window, as determined by NOAA/GOES X-ray sensor (XRS) data, with possible labels being X, M, C, B, or N (with N denoting flare-quiet, or A-class, instances).

The dataset is divided into five temporally segmented partitions, each containing approximately the same number of X- and M-class samples. Given that X- and M-class flares are of more interest, we group the samples of those two classes together as MX, while the remaining classes (C, B, and N) are grouped as NBC. Consequently, our prediction task focuses on classifying between MX and NBC, following common practices in solar flare prediction studies [8, 28, 17, 27]. SWAN-SF is highly imbalanced, with NBC samples vastly outnumbering MX samples. Table 1 summarizes the imbalance ratio and the number of samples for each class across different partition. SWAN-SF offers a realistic and challenging benchmark for solar flare prediction and provides valuable insights into the effects of class imbalance and rare event prediction in space weather analytics.

To mitigate the curse of dimensionality, we focus

our analysis on four key physical parameters: `TOTUSJH`, `TOTBSQ`, `ABSNJZH`, and `SAVNCPP`, following the approach described in [8].

3.2 GAN Architecture and Training. The GAN framework comprises two core components: a discriminator and a generator, both tailored to process temporal solar flare data. While the generator participates in adversarial training, its role is restricted to improving the discriminator's feature learning; only the discriminator is retained for downstream classification tasks. The architecture of the GAN and training pipeline for our experiments are shown in Figure 1.

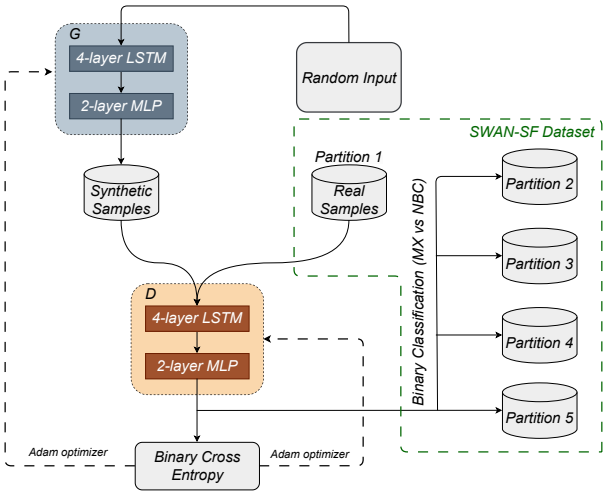

Figure 1: Schematic diagram of our GAN training and testing pipeline. The generator (G), composed of a 4-layer LSTM and a 2-layer MLP, generates synthetic samples from random inputs, while the discriminator (D), with a similar architecture, distinguishes synthetic samples from real samples (i.e., training data from Partition 1 of SWAN-SF) and performs binary classification (MX vs. NBC) on four testing partitions. Both networks are optimized using the Adam optimizer [11] with binary cross entropy loss.

3.2.1 Discriminator Architecture. Following [8], the discriminator is implemented as a multi-layer LSTM network that processes input sequences of length seq_len and feature dimension n_feat. Rather than using outputs from every time step, the network extracts the final hidden state from the LSTM as a compact representation of the entire sequence. This representation is then passed through a fully connected network composed of two linear layers with a rectified linear unit (ReLU) activation function [22] between them. The final layer applies a Sigmoid activation [24], mapping the output to a probability in the range [0, 1] to indicate the likelihood that an input sequence is real.

3.2.2 Generator Architecture. The generator is tasked with producing synthetic time-series data that closely resemble the real data distribution. It also employs an LSTM network to model the temporal dynamics of an input latent noise sequence. The output from the LSTM is fed into a time-distributed multilayer perceptron (MLP) that is applied at every time step. This MLP comprises two linear layers: the first expands the hidden representation using a ReLU activation, and the second projects the features back to the original data dimensionality. A Sigmoid activation in the output layer ensures that the generated sequences are normalized to the [0, 1] range, consistent with our preprocessed training data.

3.2.3 Weight Initialization. To ensure stable training and effective gradient propagation, we employ a custom weight initialization strategy for all linear layers in both the generator and discriminator networks. We initialize these layers with a normal distribution characterized by a mean of 0 and a standard deviation of 0.02, and set any biases to zero.

3.2.4 Leveraging the Discriminator for Binary Classification. In our study, in addition to its traditional role in distinguishing between real and generated samples during GAN training, we leverage the discriminator as a binary classifier to differentiate rare events (MX) from frequent events (NBC). The adversarial training process compels the discriminator to learn rich feature representations that capture the underlying data distributions, effectively refining its decision boundaries. By repurposing the discriminator in this manner, we directly assess its classification performance without the need for additional fine-tuning. This approach provides valuable insights into how different training strategies (as described in Section 4.1) affect model robustness and predictive accuracy in rare event scenarios.

3.3 Evaluation Metrics. Given the inherent imbalance in our testing datasets, conventional accuracy metrics can be misleading. To accurately assess the model's performance for solar flare prediction in imbalanced classification contexts, we adopt two skill scores widely used in space weather community: the True Skill Statistic (TSS) and the updated Heidke Skill Score (HSS2).

TSS quantifies the model's ability to distinguish between events and non-events. It is computed as:

$$TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN} \tag{3.1}$$

where TP, FN, FP, and TN represent the number of true positives, false negatives, false positives, and true negatives, respectively. TSS ranges from -1 to 1, with a value of 1 indicating perfect discrimination and a value of 0 indicating no skill beyond random chance.

HSS2 evaluates the forecast performance relative to random chance, accounting for both correct predictions and false alarms. It is defined as:

$$HSS2 = \frac{2 \times (TP \cdot TN - FN \cdot FP)}{P(FN + TN) + N(TP + FP)} \tag{3.2}$$

where $P = TP+FN$ and $N = FP+TN$. HSS2 provides a balanced measure by considering the imbalanced distribution of classes, ensuring that improvements in rare event detection are not overshadowed by the abundance of non-event instances. HSS2 also ranges from -1 to 1, where 1 means a perfect model and 0 means no skill beyond random guess.

However, it can be difficult to determine better performance using multiple metrics. To quantitatively compare models using these two metrics, we employ a measurement named the Distance to the Perfect (DtP) [28], which computes the Euclidean distance between a model's performance and the ideal point (TSS = 1, HSS2 = 1) on a TSS-HSS2 plot:

$$DtP = \sqrt{(1 - TSS)^2 + (1 - HSS2)^2} \tag{3.3}$$

DtP ranges from 0 to $2\sqrt{2}$, where 0 represents perfect performance, and $2\sqrt{2}$ indicates a model that predicts the exact opposite of the ground truth.

It is important to note that TSS and HSS2 are weighted equally in DtP calculation. However, since different operational settings may prioritize these metrics differently, we employ DtP solely for comparative analysis rather than as an absolute performance measure.

4 Experiments and Results.

4.1 Experiment Settings. In our experiments, Partition 1 is used as the training dataset, while Partitions 2 through 5 serve as testing datasets (shown in 1). We conduct three experiments with different training strategies for our investigations. Each strategy defines how the GAN's discriminator and generator are trained on specific subsets of the imbalanced dataset, isolating the effects of class variability and data availability from each other on feature learning and classification performance:

- **Experiment A: MX-Only Training.** In Experiment A, the GAN is trained solely on the minority class (MX) with 1,254 samples. During the training, the generator learns to generate data that resembles MX samples, and the discriminator learns to identify authentic MX samples, capturing the underlying patterns of MX data. This experiment evaluates the impact of rare-event specialization with limited training data, as the discriminator is employed directly as a binary classifier to identify the MX samples from the testing samples.

- **Experiment B: NBC-Only Training.** In Experiment B, the GAN is trained exclusively on the majority class (NBC), which consists of 72,238 samples. In contrast to Experiment A, the generator reconstructs NBC samples, while the discriminator learns to distinguish real NBC samples from their reconstructions. This approach forces the discriminator to focus on the characteristic patterns associated with weak-flare (and non-flaring) behavior. After adversarial training, the discriminator is directly used as a binary classifier by identifying those samples that are more likely to be NBC in the testing data.

- **Experiment C: Undersampled NBC Training.** In Experiment C, a subset of NBC samples is created by randomly undersampling (without replacement) the majority class to match the number of MX samples. The GAN is then trained on the undersampled NBC data, ensuring equal data availability across classes compared to Experiment A and allowing us to assess the influence of data quantity relative to Experiment B.

To ensure reliability, we run each experiment for multiple times. Both Experiment A and Experiment B are run with ten different weight initializations to demonstrate the robustness of the model under varying machine learning regimes. In Experiment C, since variability arises from both weight initialization and the random undersampling process, we conduct three weight initialization runs, each repeated across four different random undersampling configurations. For each run, we select the best-performing model that achieves the best average DtP across the four testing partitions from SWAN-SF. We then compute the mean and standard deviation of the evaluation metrics from these best-performing models across multiple runs, yielding our final measures of performance and robustness for solar flare prediction in each experiment.

For Experiment B and Experiment C, we train the GAN using the Adam [11] optimizer with a learning rate of 0.0002 and a batch size of 32 for 500 epochs. For Experiment A, we use the same configurations except for the batch size, which is set to 256 due to the much larger training dataset.

Table 2: Performance of the discriminator on different testing partitions in each experiment.

Experiment	Metric	Partition 2	Partition 3	Partition 4	Partition 5	Average
A (MX-Only)	TSS	0.62 ± 0.03	0.65 ± 0.01	0.80 ± 0.02	0.72 ± 0.01	0.70 ± 0.07
	HSS2	0.32 ± 0.01	0.33 ± 0.01	0.40 ± 0.01	0.35 ± 0.02	0.35 ± 0.03
	DtP	$\mathbf{0.78 \pm 0.01}$	$\mathbf{0.76 \pm 0.01}$	$\mathbf{0.63 \pm 0.01}$	$\mathbf{0.70 \pm 0.02}$	$\mathbf{0.72 \pm 0.06}$
B (NBC-Only)	TSS	0.65 ± 0.13	0.65 ± 0.08	0.79 ± 0.06	0.72 ± 0.09	0.70 ± 0.11
	HSS2	0.29 ± 0.04	0.30 ± 0.02	0.37 ± 0.06	0.31 ± 0.04	0.32 ± 0.05
	DtP	0.80 ± 0.05	0.78 ± 0.03	0.67 ± 0.04	0.75 ± 0.02	0.75 ± 0.06
C (Undersampled NBC)	TSS	0.66 ± 0.08	0.64 ± 0.05	0.79 ± 0.02	0.72 ± 0.05	0.70 ± 0.08
	HSS2	0.30 ± 0.03	0.32 ± 0.01	0.39 ± 0.04	0.34 ± 0.03	0.34 ± 0.04
	DtP	0.78 ± 0.03	0.77 ± 0.01	0.65 ± 0.03	0.72 ± 0.02	0.73 ± 0.06

4.2 Results and Discussion. Table 2 presents the performance results for the different experiments across various testing partitions. Each cell displays the *mean* $\pm$ *std* of the corresponding metrics over the multiple runs described in Section 4.1.

Among the three training strategies, Experiment A (MX-only training) achieved the lowest average DtP across all testing partitions and the smallest standard deviation over multiple runs. This suggests not only a more precise placement of the decision boundary but also greater robustness. Even with limited data, training exclusively on MX samples enabled the GAN's discriminator to learn highly discriminative features that more effectively separate MX from NBC samples. Although the majority class (NBC) is typically assumed to exhibit more consistent patterns, the results imply that its underlying distribution may be more complex, which may offset the potential benefits of having approximately 60 times more data.

Comparing Experiment A and Experiment C, where both classes have equal data availability, reveals that the discriminator trained solely on MX samples achieves a slightly better and more robust performance. This suggests that the difference of class variability between MX and NBC may not be as significant as presumed. One potential explanation is that NBC comprises three different sub-classes (C, B, and N), introducing additional variability into the majority class.

The comparison between Experiment B and Experiment C reveals that training with the entire NBC class does not necessarily enhance solar flare prediction. In fact, the discriminator trained on a subset of NBC samples performs slightly better and more stably. Using all available NBC samples may expose the model to a wide array of patterns, many of which do not contribute to distinguishing NBC from MX. This may lead to overfitting on non-discriminative features and hinder the development of a robust decision boundary. Conversely, undersampling NBC to match the number of MX samples encourages the model to focus on the most discriminative features, reducing the influence of redundant or irrelevant patterns and resulting in lower performance variability.

Furthermore, comparing Experiment A and Experiment B shows that the discriminator trained exclusively on MX samples performs more robustly despite having access to significantly less data. This may be because the inclusion of three NBC sub-classes introduces additional variability that diminishes the benefits from a larger data availability. Another possible reason could be that the GAN model lacks the complexity needed to fully exploit the extensive NBC data.

Overall, the experimental results indicate that the distribution of MX samples appears to be more homogeneous than that of NBC samples in Partition 1 from SWAN-SF. This finding contradicts our initial assumption that rare-event data would be more variable than frequent-event data in the context of solar flare prediction due to their rarity. Additionally, the results suggest that data availability does not necessarily outweigh class variety when training machine learning models for solar flare prediction.

5 Conclusion and Future Work.

In this study, we investigated the impact of class distribution on the performance of a GAN-based solar flare prediction model using the SWAN-SF dataset. We explored three training strategies, including NBC-only training, MX-only training, and undersampled NBC training, to isolate the effects of data imbalance on feature learning and decision-boundary precision. Our results indicate that the discriminator trained solely on MX samples achieves the lowest DtP values and exhibits robust performance across testing partitions. This

finding suggests that the MX class, despite being a rare event, may possess a more homogeneous distribution than initially assumed, challenging the common belief that rare-event data is usually inherently more variable than frequent-event data. Moreover, our experiments demonstrate that simply having more data (as in the case of NBC-only training) does not necessarily lead to better predictive performance, highlighting the critical role of class variety in rare-event prediction.

These insights emphasize the need for a careful balance between data quantity and quality when training machine learning models on imbalanced datasets. In our case, focusing on a smaller, more homogeneous dataset led to a more effective and robust discriminator, which is crucial for distinguishing between rare and frequent events.

For future work, several promising directions can be pursued:

- **Advanced Data Balancing Techniques:** Investigate data augmentation, oversampling, or synthetic data generation methods to further enhance the representation of minority classes.

- **Model Architecture Enhancements:** Explore alternative GAN architectures or hybrid models that integrate adversarial and supervised learning components to improve overall predictive performance.

- **Sub-class Analysis:** Conduct a deeper analysis of the sub-class distributions within the majority class (NBC) to understand their contribution to model instability and to devise strategies for mitigating their adverse effects.

By addressing these avenues, future research can further refine the predictive capabilities of machine learning models for rare-event scenarios and contribute to more accurate and reliable solar flare forecasting.

References

[1] Azim Ahmadzadeh, Berkay Aydin, Manolis K Georgoulis, Dustin J Kempton, Sushant S Mahajan, and Rafal A Angryk. How to train your flare prediction model: Revisiting robust sampling of rare events. *The Astrophysical Journal Supplement Series*, 254(2):23, 2021.

[2] Giuseppina Andresini, Annalisa Appice, Luca De Rose, and Donato Malerba. Gan augmentation to deal with imbalance in imaging-based intrusion detection. *Future Generation Computer Systems*, 123:108–127, 2021.

[3] Rafal A Angryk, Petrus C Martens, Berkay Aydin, Dustin Kempton, Sushant S Mahajan, Sunitha Basodi, Azim Ahmadzadeh, Xumin Cai, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi, et al. Multivariate time series dataset for space weather data analytics. *Scientific data*, 7(1):227, 2020.

[4] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[5] Monica G Bobra and Sebastien Couvidat. Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, 798(2):135, 2015.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[7] Yang Chen, Dustin J Kempton, Azim Ahmadzadeh, and Rafal A Angryk. Towards synthetic multivariate time series generation for flare forecasting. In *Artificial Intelligence and Soft Computing: 20th International Conference, ICAISC 2021, Virtual Event, June 21–23, 2021, Proceedings, Part I 20*, pages 296–307. Springer, 2021.

[8] Yang Chen, Dustin J Kempton, Azim Ahmadzadeh, Junzhi Wen, Anli Ji, and Rafal A Angryk. Cgan-based synthetic multivariate time-series generation: a solution to data scarcity in solar flare forecasting. *Neural Computing and Applications*, 34(16):13339–13353, 2022.

[9] Yang Chen, Dustin J Kempton, and Rafal A Angryk. Examining effects of class imbalance on conditional gan training. In *International Conference on Artificial Intelligence and Soft Computing*, pages 475–486. Springer, 2023.

[10] Yang Chen, Dustin J Kempton, and Rafal A Angryk. Ffad: A novel metric for assessing generated time series data utilizing fourier transform and auto-encoder. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 129–139. Springer, 2024.

[11] Kingma Diederik. Adam: A method for stochastic optimization. *(No Title)*, 2014.

[12] Yutong Gao, Vince D Calhoun, and Robyn L Miller. Transient intervals of significantly different whole brain connectivity predict recovery vs. progression from mild cognitive impairment: new insights from interpretable lstm classifiers. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4645–4648. IEEE, 2022.

[13] Yutong Gao, Vince D Calhoun, and Robyn L Miller. Generative forecasting of brain activity enhances alzheimer's classification and interpretation. *arXiv preprint arXiv:2410.23515*, 2024.

[14] Manolis K Georgoulis and David M Rust. Quantitative forecasting of major solar flares. *The Astrophysical*

Journal, 661(1):L109, 2007.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[17] Anli Ji, Junzhi Wen, Rafal Angryk, and Berkay Aydin. Solar flare forecasting with deep learning-based time series classifiers. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2907–2913. IEEE, 2022.

[18] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks*, pages 703–716. Springer, 2019.

[19] Hao Liu, Chang Liu, Jason TL Wang, and Haimin Wang. Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, 877(2):121, 2019.

[20] Ruizhe Ma, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi, and Rafal A Angryk. Solar flare prediction using multivariate time series decision trees. In *2017 IEEE international conference on big data (big data)*, pages 2569–2578. IEEE, 2017.

[21] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[22] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[23] Chetraj Pandey, Temitope Adeyeha, Jinsu Hong, Rafal A Angryk, and Berkay Aydin. Advancing solar flare prediction using deep learning with active region patches. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 50–65. Springer, 2024.

[24] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[25] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, 2017.

[26] Fei Wang, Zhao Zhen, Bo Wang, and Zengqiang Mi. Comparative study on knn and svm based weather classification models for day ahead short term solar pv power forecasting. *Applied Sciences*, 8(1):28, 2017.

[27] Junzhi Wen, Azim Ahmadzadeh, Manolis K Georgoulis, Viacheslav M. Sadykov, and Rafal A Angryk. Outlier detection and removal in multivariate time series for a more robust machine learning-based solar flare prediction. *The Astrophysical Journal Supplement Series*, 2025. In press.

[28] Junzhi Wen and Rafal A. Angryk. Class-based time series data augmentation to mitigate extreme class imbalance for solar flare prediction. In Leszek Rutkowski, Rafał Scherer, Marcin Korytkowski, Witold Pedrycz, Ryszard Tadeusiewicz, and Jacek M. Zurada, editors, *Artificial Intelligence and Soft Computing*, pages 362–375, Cham, 2025. Springer Nature Switzerland.

[29] Junzhi Wen, Md Reazul Islam, Azim Ahmadzadeh, and Rafal A Angryk. Improving solar flare prediction by time series outlier detection. In *International Conference on Artificial Intelligence and Soft Computing*, pages 152–164. Springer, 2022.

[30] MS Wheatland. A bayesian approach to solar flare prediction. *The Astrophysical Journal*, 609(2):1134, 2004.

[31] MS Wheatland. A statistical solar flare forecast method. *Space Weather*, 3(7), 2005.