ARENA: Active Region attENtion Aggregation for interpretable solar flare prediction

Luis Bauer
MPI for Intelligent Systems, Tübingen
University of Tübingen

Annalena Kofler*
MPI for Intelligent Systems, Tübingen
MPI for Gravitational Physics, Potsdam

Andrés Asensio Ramos
Instituto de Astrofísica de Canarias
Universidad de La Laguna

V. Aparna
New Mexico State University

Valentín Martínez Pillet
Instituto de Astrofísica de Canarias
Universidad de La Laguna

Bernhard Schölkopf
MPI for Intelligent Systems, Tübingen
ELLIS Institute Tübingen

Abstract

Predicting solar flares is a challenging space-weather forecasting problem with potentially significant impacts on Earth's upper atmosphere, including disruptions to radio communications. We propose **Active Region attENtion Aggregation (ARENA)**, a neural network architecture that predicts solar flares from a variable number of active-region cutouts while learning to identify the region most relevant to the prediction. ARENA is trained using multiple-instance learning under weak supervision, requiring only a flare/no-flare label for the observation and no labels specifying the flare origin. The learned attention mechanism therefore provides a means of localizing the active region driving the prediction without explicit origin annotations. ARENA achieves a heidke skill score of 0.45 which is comparable to other flare prediction approaches. These results demonstrate that active-region-based aggregation can provide an efficient and interpretable approach to solar flare prediction. The code is available at https://github.com/LuisBauer22/ARENA.

1 Introduction

Solar flare forecasting. Solar flares are transient releases of magnetic energy in the solar atmosphere that produce intense electromagnetic radiation. These events are associated with active regions (ARs), localized concentrations of strong and complex magnetic fields on the solar surface [1]. Radiation from the most energetic flares can perturb Earth's ionosphere and affect radio communication and space-based technologies, making reliable flare forecasting an important problem in space-weather research. A central challenge in solar flare forecasting is the spatial scale at which the problem is formulated: A flare is observed as a global event, but its magnetic origin is localized to an AR. Consequently, two related questions arise: *Will a flare occur within 24 h, and which AR is responsible for it?* The first question can be addressed using event-level flare catalogs, whereas the second requires associating individual flares with their source regions. Most existing data-driven forecasting approaches focus on the former, using either full-disk observations [2–4] or individual ARs [5–8] as

*Corresponding author: annalena.kofler@tuebingen.mpg.de

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Representations for the Physical Sciences.

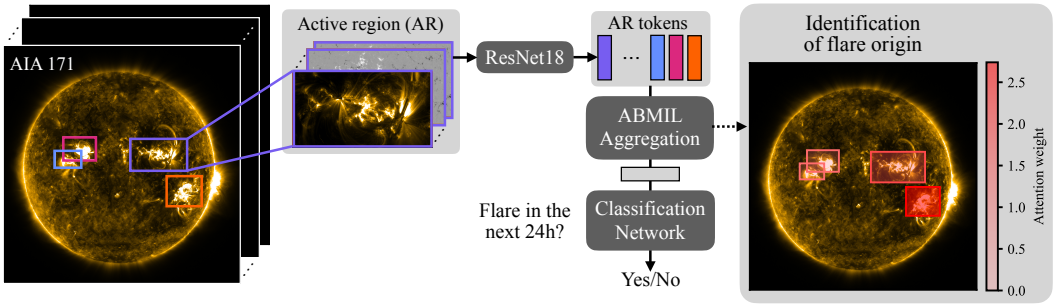


Figure 1: Overview of the ARENA architecture. Each AR across all channels is extracted, rescaled, and provided to a ResNet-18 which compresses the image into a token vector. For one time point, all AR tokens are concatenated and passed through an ABMIL aggregator, resulting in a single summary vector. This vector is processed by a classification head which predicts whether a flare is happening in the next 24 h. At inference, the attention weight can be employed to identify the flare-originating AR.

input. In contrast, combining ARs present in an observation provides a natural representation of the solar disk as a variable-sized set of physically meaningful regions.

Multiple-instance learning for active region identification. Representations combining ARs introduce a mismatch between the available data and the desired labels. For each observation, it is straightforward to obtain a global label indicating whether a flare occurs within the subsequent 24 h. Determining which of the ARs present at the observation time produces the subsequent flare, however, requires additional event-to-region association based on expert knowledge. Such origin annotations are consequently substantially more difficult to obtain and are available only for limited datasets [9, 10]. We therefore seek to learn from the readily available global flare labels while retaining the ability to identify the AR driving the prediction. This setting corresponds naturally to *multiple-instance learning* (MIL [11]). In MIL, a bag contains multiple instances but is associated only with a bag-level label. This paradigm has been particularly successful in computational pathology, where a high-resolution whole-slide image may be labeled according to the presence of a disease, such as cancer, while identifying the diagnostically relevant tissue regions requires costly expert annotation [12–15]. We formulate solar flare forecasting analogously: An observation constitutes a bag, its associated AR cutouts are the instances, and the flare/no-flare label is assigned to the bag. The number of ARs can vary between observations, requiring an aggregation mechanism that is invariant to the number and ordering of instances. Importantly, information about the AR associated with a flare is not available during training.

Contributions. (1) We propose **ARENA (Active Region attENtion Aggregation)**, an attention-based neural network architecture that represents solar observations as variable-sized sets of physically motivated active region instances, illustrated in Figure 1. (2) We formulate solar flare forecasting as a MIL problem, using only observation-level flare labels while treating the associated active regions as instances. This enables training without explicit labels for the flare-producing region. (3) We show that the learned attention weights can identify the AR associated with the observed flare, despite the absence of origin annotations during training. We show that the ARENA model provides an interpretable localization signal while maintaining comparable flare forecasting performance.

2 ARENA: Active Region attENtion Aggregation

MIL problem formulation. We formulate solar flare forecasting as a MIL problem where each observation at time t is represented as a bag $\mathcal{B}(t) = \{\mathbf{x}_1, ..., \mathbf{x}_{N_t}\}$ containing the active regions identified in the corresponding AR catalog. Here, $\mathbf{x}_i$ denotes the image cutout of the ith AR at time t and N_t denotes the number of ARs present at that timestamp. Consequently, the number of instances varies between observations. Observations containing no ARs are excluded from the dataset; the distribution of N_t is reported in Appendix A. This AR-based data representation focuses the model on physically relevant regions and avoids processing the substantially larger full-disk images, significantly reducing memory and disk requirements.

Each bag is assigned a binary label

$$y_t = \begin{cases} 1 & \exists t_{\text{flare}} : \; t < t_{\text{flare}} \leq t + 24\,\text{h} \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{1}$$

if there is at least one M- and X-class flare within the subsequent 24 hours. During training, only the bag-level label y_t is provided. For a positive bag, the active region responsible for the subsequent flare is not annotated; instead, we assume that the flare-producing AR is among the ARs present at time t. Consequently, the model must learn which ARs are relevant to the bag-level prediction from the global flare label alone. This constitutes weak supervision at the AR level and motivates an attention-based aggregation mechanism capable of handling a variable number of AR instances.

ARENA architecture. ARENA encodes each AR independently and subsequently aggregates the resulting representations into a single bag-level representation. Given an observation $\mathcal{B}(t)$ as illustrated in Figure 1, each AR cutout is mapped to an embedding $\mathbf{h_i} = f_\theta(\mathbf{x_i})$, where f_θ is a ResNet-18 [16] image encoder. For the main ARENA model, we use attention-based multiple-instance learning (ABMIL [12]) to aggregate the AR embeddings. Specifically, the attention weights are computed as

$$\alpha_i = \frac{\exp\left[a_\phi(\mathbf{h_i})\right]}{\sum_{j=1}^{N_t} \exp\left[a_\phi(\mathbf{h_j})\right]} \,, \tag{2}$$

where a_ϕ is a learned attention function. The resulting bag representation is given by the attention-weighted sum $\mathbf{z} = \sum_{i=1}^{N_t} \alpha_i \mathbf{h_i}$. The exact gated attention formulation and its architectural details are provided in Appendix B. A classification head then maps $\mathbf{z}$ to the probability of an M+ flare within the following 24 h. Since ARENA performs the aggregation in the learned feature space, it can learn both which AR representations are relevant and which features within these representations contribute to the bag-level prediction.

3 Experimental Setup

Solar data. We use solar observations from 2010–2024 at a cadence of two hours, with the active regions identified from the corresponding AR catalog and represented by image cutouts [17]. Flare labels are obtained from the NOAA Geostationary Operational Environmental Satellite X-ray sensor flare (GOES XRS) catalog [18]. The data from 2010–2016 are used for training and validation, while 2017–2024 are held out as an independent test set. Within the development period, we perform five temporal validation folds, each using two months of each year for validation and a one-month temporal buffer between the training and validation periods to avoid temporal temporal leakage. For evaluating AR source identification, we use the subset of the flare dataset containing annotations of the AR associated with each flare [9] corresponding to a set of flares in 2017. Further details on data preprocessing and dataset composition are provided in Appendix A.

Training. Each model is optimized using AdamW [19] with a cross-entropy loss. All five models are subsequently evaluated on the same held-out test period from 2017–2024. Further details on training hyperparameters are provided in Appendix C.

Evaluation. We evaluate ARENA along two complementary dimensions: Flare forecasting and AR source identification. For flare forecasting, we use the Heidke Skill Score (HSS) as the primary metric across the validation and temporally held-out test sets. The HSS measures the improvement of a binary classifier over a random forecast with optimal class frequencies, see Appendix D. To assess whether the weakly supervised aggregation limits predictive performance, we additionally compare ARENA with an instance-level classifier using the same AR encoder but trained with explicit flare-AR annotations [9]. For this comparison, we report additional classification metrics to characterize differences between direct AR-level and bag-level prediction. Details of the evaluation protocol and baseline comparisons are provided in Appendix B and Appendix E, respectively.
For AR source identification, we rank ARs within each positive bag according to their attention weights and quantify source-AR localization using top-k accuracy (Acc@k). This measures the frac-

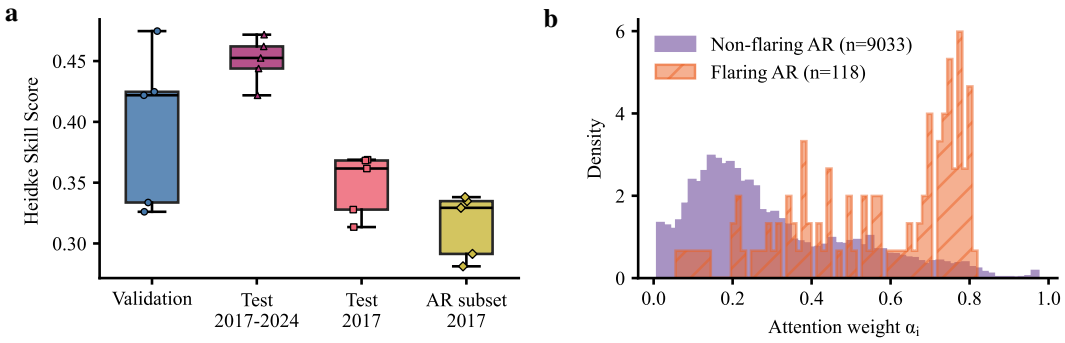


Figure 2: **(a)** Heidke Skill Score (HSS) across the five temporal cross-validation folds, evaluated on the corresponding validation sets and on three test sets: the full 2017–2024 test period, the 2017 subset, and the 2017 subset containing flares with confirmed AR origins. **(b)** Density-normalized distribution of attention weights assigned to flare-producing and non-producing ARs for the 2017 subset with confirmed AR origins, using the model with the highest validation HSS among the five cross-validation models.

tion of observations for which the annotated flare-producing AR is among the k highest-attention ARs:

$$\mathrm{Acc@}k = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\mathrm{rank}_n(i_n^*) \leq k\right] \ , \tag{3}$$

where i_n^* denotes the annotated flare-producing AR and $\mathrm{rank}_n(i_n^*)$ its rank according to the attention weights. We additionally analyze the distribution of attention weights assigned to flare-producing and non-producing ARs. Further details including explicit metric formulae and comparable source identification works are provided in Appendix D.

4 Results

Flare forecasting on future data. We first evaluate ARENA on the temporally held-out test set from 2017–2024. Figure 2a shows the HSS obtained by the five models trained on different temporal cross-validation folds, evaluated on both their respective validation sets and the common test sets. The models exhibit varying performance across the validation folds which could be related to varying class balances across the folds (see Appendix E). Since the 2017–2024 test set contains significantly more data than the validation folds, we consider it to provide a more robust result. This test set shows a median HSS of 0.45, slightly higher than the median HSS on the validation data (HSS$= 0.42$). The AR identification for flares can only be evaluated on a subset of the 2017 data due to limited flare-AR labels [9]. Therefore we show the HSS for the full 2017 data as well as the specific subset where AR region labels are available. ARENA performs comparably to the AR-level supervised baseline (Table 1), suggesting that the attention-based aggregation mechanism introduces little performance degradation and that further improvements may instead be limited by the AR encoder.

Table 1: Comparison of per-AR instance flare level prediction vs bag level prediction on the same subset of 2017. We report the median metrics across the five validation folds.

	TSS↑	HSS↑	F1↑	Precision↑	TPR↑	TNR↑
Per AR (instance level)	**0.85**	0.17	0.19	0.10	**0.96**	**0.90**
ARENA (bag level)	0.79	**0.33**	**0.37**	**0.23**	0.92	0.88

AR source identification. We next assess whether the learned attention weights identify the AR associated with a subsequent flare. Excluding observations containing only one AR for which source localization is trivial, ARENA achieves an Acc@1 of $87.1 \pm 0.7\%$ and an Acc@2 of $100.0 \pm 0.0\%$

across the five models. Figure 2b shows the corresponding distribution of attention weights for flare-producing and non-producing ARs for our best run (validation HSS$= 0.47$). Flare-producing ARs receive systematically higher attention weights, consistent with the quantitative localization results.

5 Conclusion

We introduced **ARENA**, an attention-based multiple-instance learning architecture for interpretable solar flare forecasting. Using only bag-level flare labels and AR cutouts, ARENA achieves a median HSS of 0.45 on the temporally held-out 2017–2024 test set and reaches a performance comparable to a model trained on flare-AR labels, while providing an intrinsic AR-level importance signal. Despite never observing flare-origin annotations during training, ARENA assigns the highest attention weight to the flare-producing AR in $87.1 \pm 0.7\%$ of cases. These results demonstrate that weak supervision can recover meaningful AR-level associations directly through the model's attention mechanism. Future work will extend this framework to other solar eruptive phenomena, different attention mechanisms and automated flare-AR attribution.

Acknowledgments and Disclosure of Funding

The authors thank Vincent Berenz for great support in storing and handling large amounts of solar images and Hsiao-Ru Pan for the thoughtful discussions on the computer vision approaches. AK thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for support. The computational work for this paper was performed on the Atlas cluster at the Max Planck Institute for Intelligent Systems. AAR and VMP acknowledge financial support from the Agencia Estatal de Investigación (AEI) through the Severo Ochoa Centre of Excellence accreditation awarded to the Instituto de Astrofísica de Canarias, grant CEX2025-001609-S, funded by MICIU/AEI/10.13039/501100011033.

References

[1] L. Fletcher, B. R. Dennis, H. S. Hudson, S. Krucker, K. Phillips et al. An observational overview of solar flares. *Space Sci Rev*, 159(19), 2011. doi: 10.1007/s11214-010-9701-8. URL https://link.springer.com/article/10.1007/s11214-010-9701-8.

[2] C. Pandey, R. A. Angryk, and B. Aydin. Explaining full-disk deep learning model for solar flare prediction using attribution methods. In *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track*, pages 72–89, Cham, 2023. Springer Nature Switzerland. doi: 10.1007/978-3-031-43430-3_5. URL https://link.springer.com/chapter/10.1007/978-3-031-43430-3_5.

[3] S. Roy, J. Schmude, R. Lal, V. Gaur, M. Freitag et al. Surya: Foundation model for heliophysics, 2025. URL https://arxiv.org/abs/2508.14112.

[4] P. Yan, X. Li, Y. Zheng, L. Dong, S. Yan et al. A real-time solar flare forecasting system with deep learning methods. *Astrophysics and Space Science*, 369(10):110, 2024. doi: 10.1007/s10509-024-04374-8. URL https://link.springer.com/article/10.1007/s10509-024-04374-8.

[5] M. Li, Y. Cui, B. Luo, J. Wang, and X. Wang. Deep neural networks of solar flare forecasting for complex active regions. *Frontiers in Astronomy and Space Sciences*, Volume 10 - 2023, 2023. doi: 10.3389/fspas.2023.1177550. URL https://www.frontiersin.org/journals/astronomy-and-space-sciences/articles/10.3389/fspas.2023.1177550.

[6] Z. Sun, M. G. Bobra, X. Wang, Y. Wang, H. Sun et al. Predicting solar flares using cnn and lstm on two solar cycles of active region data. *The Astrophysical Journal*, 931(2):163, 6 2022. doi: 10.3847/1538-4357/ac64a6. URL https://iopscience.iop.org/article/10.3847/1538-4357/ac64a6.

[7] R. Tang, W. Liao, Z. Chen, X. Zeng, J.-s. Wang et al. Solar flare prediction based on the fusion of multiple deep-learning models. *The Astrophysical Journal Supplement Series*, 257(2):50, 12

2021. doi: 10.3847/1538-4365/ac249e. URL https://iopscience.iop.org/article/10.3847/1538-4365/ac249e.

[8] X. Huang, H. Wang, L. Xu, J. Liu, R. Li et al. Deep learning based solar flare forecasting model. I. results for line-of-sight magnetograms. *The Astrophysical Journal*, 856(1):7, 3 2018. doi: 10.3847/1538-4357/aaae00. URL https://iopscience.iop.org/article/10.3847/1538-4357/aaae00.

[9] R. A. Angryk, P. C. Martens, B. Aydin, D. Kempton, S. S. Mahajan et al. Multivariate time series dataset for space weather data analytics. *Sci Data*, 7(227), 2020. doi: 10.1038/s41597-020-0548-x. URL https://www.nature.com/articles/s41597-020-0548-x.

[10] L. E. Boucheron, T. Vincent, J. A. Grajeda, and E. Wuest. Solar active region magnetogram image dataset for studies of space weather. *Scientific Data*, 10(1):825, 2023. doi: 10.1038/s41597-023-02628-8. URL https://www.nature.com/articles/s41597-023-02628-8.

[11] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018. doi: 10.1016/j.patcog.2017.10.009. URL https://www.sciencedirect.com/science/article/pii/S0031320317304065.

[12] M. Ilse, J. Tomczak, and M. Welling. Attention-based deep multiple instance learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. doi: 10.48550/arXiv.1802.04712. URL https://proceedings.mlr.press/v80/ilse18a.html.

[13] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021. doi: 10.48550/arXiv.2106.00908. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/10c272d06794d3e5785d5e7c5356e9ff-Paper.pdf.

[14] Y. Cui, Z. Liu, X. Liu, X. Liu, C. Wang et al. Bayes-MIL: A new probabilistic perspective on attention-based multiple instance learning for whole slide images. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_geIwiOyUhZ.

[15] W. Zhu, X. Chen, P. Qiu, A. Sotiras, A. Razi et al. DGR-MIL: Exploring diverse global representation in multiple instance learning for whole slide image classification. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision – ECCV 2024*, pages 333–351, Cham, 2025. Springer Nature Switzerland. doi: 0.1007/978-3-031-72920-1_19. URL https://link.springer.com/chapter/10.1007/978-3-031-72920-1_19.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. doi: 10.1109/CVPR.2016.90. URL https://api.semanticscholar.org/CorpusID:206594692.

[17] M. Bobra, X. Sun, T. Hoeksema, M. Turmon, Y. Liu et al. The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPs – Space-weather HMI Active Region Patches. *Solar Physics*, 04 2014. doi: 10.1007/s11207-014-0529-3. URL https://link.springer.com/article/10.1007/s11207-014-0529-3.

[18] J. L. Machol, S. M. Codrescu, C. Peck, et al. GOES-R Series X-Ray Sensor (XRS): 2. On-orbit measurements and calibrations. *Journal of Geophysical Research: Space Physics*, 02 2026. doi: 10.22541/essoar.177092672.25576917/v1. URL https://essopenarchive.org/doi/full/10.22541/essoar.177092672.25576917/v1.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. doi: 10.48550/arXiv.1412.6980. URL https://api.semanticscholar.org/CorpusID:6628106.

[20] J. Hoeksema, Y. Liu, K. Hayashi, et al. The helioseismic and magnetic imager (HMI) vector magnetic field pipeline: Overview and performance. *Solar Physics*, 289(9):3483–3530, 2014. doi: 10.1007/s11207-014-0516-8. URL https://link.springer.com/article/10.1007/s11207-014-0516-8.

[21] C. Pandey, T. Adeyeha, J. Hong, R. A. Angryk, and B. Aydin. Advancing Solar Flare Prediction Using Deep Learning with Active Region Patches. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 50–65, Cham, 2024. Springer Nature Switzerland. doi: 10.1007/978-3-031-70381-2_4. URL https://link.springer.com/chapter/10.1007/978-3-031-70381-2_4.

[22] S. Guastavino, F. Marchetti, F. Benvenuto, C. Campi, and M. Piana. Operational solar flare forecasting via video-based deep learning. *Frontiers in Astronomy and Space Sciences*, Volume 9 - 2022, 2023. doi: 10.3389/fspas.2022.1039805. URL https://www.frontiersin.org/journals/astronomy-and-space-sciences/articles/10.3389/fspas.2022.1039805.

[23] 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*, 256(2):26, September 2021. doi: 10.3847/1538-4365/ac1f1d. URL https://iopscience.iop.org/article/10.3847/1538-4365/ac1f1d.

[24] J. R. Lemen, A. M. Title, D. J. Akin, P. F. Boerner, C. Chou et al. The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):17–40, jan 2012. doi: 10.1007/s11207-011-9776-8. URL https://link.springer.com/article/10.1007/s11207-011-9776-8.

[25] G. Francisco, M. Berretti, S. Chierichini, R. Mugatwala, J. M. Fernandes et al. Limits of Solar flare Forecasting Models and New Deep Learning Approach. *The Astrophysical Journal*, 985 (1):108, 5 2025. doi: 10.3847/1538-4357/adc56d. URL https://iopscience.iop.org/article/10.3847/1538-4357/adc56d.

[26] K. Dissauer, K. D. Leka, and E. L. Wagner. Properties of flare-imminent versus flare-quiet Active Regions from the Chromosphere through the Corona. I. Introduction of the AIA Active Region Patches (AARPs). *The Astrophysical Journal*, 942(2):83, Jan. 2023. doi: 10.3847/1538-4357/ac9c06. URL https://iopscience.iop.org/article/10.3847/1538-4357/ac9c06.

[27] J. Schou, P. H. Scherrer, R. I. Bush, R. Wachter, S. Couvidat et al. Design and Ground Calibration of the Helioseismic and Magnetic Imager (HMI) Instrument on the Solar Dynamics Observatory (SDO). *Solar Physics*, 275(1-2):229–259, jan 2012. doi: 10.1007/s11207-011-9842-2. URL https://doi.org/10.1007/s11207-011-9842-2.

[28] D. Bloomfield, P. Higgins, J. Mcateer, and P. Gallagher. Toward reliable benchmarking of solar flare forecasting methods. *The Astrophysical Journal*, 747:L41, 02 2012. doi: 10.1088/2041-8205/747/2/L41. URL https://iopscience.iop.org/article/10.1088/2041-8205/747/2/L41.

[29] C. Pandey, A. Ji, R. A. Angryk, and B. Aydin. Towards interpretable solar flare prediction with attention-based deep neural networks. In *2023 IEEE Sixth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 83–90, 2023. doi: 10.1109/AIKE59827.2023.00021. URL https://ieeexplore.ieee.org/abstract/document/10381566.

[30] K. Yi, Y.-J. Moon, D. Lim, E. Park, and H. Lee. Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters. *The Astrophysical Journal*, 910(1):8, 22 March 2021. doi: 10.3847/1538-4357/abdebe. URL https://iopscience.iop.org/article/10.3847/1538-4357/abdebe.

[31] C. Huwyler and M. Melchior. Using multiple instance learning for explainable solar flare prediction. *Astronomy and Computing*, 41:100668, 2022. doi: 10.1016/j.ascom.2022.100668. URL https://www.sciencedirect.com/science/article/pii/S2213133722000828.

A Data set and preprocessing

Active region image catalog. We represent each solar observation as a variable-sized set of active regions (ARs) rather than using full-disk observations. ARs are the primary sites of solar flare activity, particularly for large and energetic flares, and therefore provide a physically motivated spatial decomposition of the solar disk. Restricting the input to AR cutouts also avoids processing the large fraction of the solar disk corresponding to quiet-Sun regions, substantially reducing memory, computational, and storage requirements while retaining the regions most relevant to flare prediction. We use the Space-weather HMI Active Region Patches (SHARPs) catalog [17, 20], which provides automatically identified and tracked AR patches derived from observations of the Helioseismic and Magnetic Imager (HMI). SHARPs are widely used for data-driven solar flare prediction [6–8, 21, 22]. The SHARPs catalog is derived from observations of the photospheric magnetic field. A magnetogram measures the magnetic field projected onto a particular direction. In a line-of-sight (LOS) magnetogram, the measured quantity is the component of the magnetic field along the line connecting the solar surface to the observer. In contrast, vector magnetograms provide information about the full magnetic-field vector, including both the LOS and transverse components. The active-region detection and tracking pipeline used to construct SHARPs operates on line-of-sight and vector magnetograms and identifies regions of enhanced magnetic activity. In particular, the detection procedure applies Gaussian-kernel filtering to activity bitmaps derived from the magnetograms [23].

Multi-channel active-region dataset. We construct a custom dataset based on the SHARPs catalog by augmenting the magnetic-field information with observations of the solar atmosphere: Extreme ultra-violet (EUV) observations provide complementary information about the morphology of the solar corona, including coronal loops and other structures associated with active regions. In particular, the Atmospheric Imaging Assembly (AIA) channels [24] at 131, 171 and 304 Å are each dominated by emission from ions with distinct characteristic formation temperatures, so that the three channels jointly sample plasma across a broad range of coronal and transition-region temperatures. This provides morphological information that is not directly captured by photospheric magnetograms alone. Each channel image encodes intensity as a function of position rather than an explicit temperature; we do not perform a differential-emission-measure inversion or otherwise reconstruct a temperature map.

Instead, by treating the co-aligned channels as separate input channels to the network, we allow the model to exploit channel-specific and cross-channel morphological differences (e.g. a structure that is prominent in one band but faint in another). These differences arise from the differing temperature sensitivity of each channel, without requiring the network to represent temperature explicitly.

Recent work has also shown that flare-forecasting models using EUV observations can outperform models based solely on magnetograms [25]. We also increase the spatial extent of the extracted patches relative to the standard SHARP cutouts, similar to Dissauer et al. [26]. This provides additional context around the active region and captures larger-scale coronal structures, such as extended magnetic loops, that may extend beyond the photospheric AR boundary. This way each AR instance can be viewed as a multi-modal physical representation of the photospheric magnetic field and coronal morphology, with an example visualized in Figure 3 and the corresponding full-disk images in Figure 4. We refer to the resulting multi-channel active-region cutouts as *AIA Active Region Patches* (AARPs) throughout the remainder of the paper.

For extracting the same solar region from AIA and vector magnetogram channels we use the multi-channel full-disk images of the Surya-dataset [3], covering 05/2010-12/2024 at a 12 min cadence. The Surya dataset is based on images from Solar Dynamics Observatory (SDO), specifically from the eight AIA channels [24] and five HMI channels [27]. The AIA instrument captures EUV and UV images at a cadence of 12 s to 24 s while the HMI instrument captures images of the magnetic field of the sun at a cadence of 12 s. We use the timestamps provided by Surya at a reduced cadence of two hours. This temporal sampling is chosen because the energy build up and large-scale evolution of the magnetic and coronal structures preceding flare events occurs predominantly on timescales of hours to days, while the native instrument cadence would result in substantial redundancy for our forecasting time horizon. For each selected timestamp, we identify the corresponding ARs and extract spatially aligned multi-channel cutouts from the HMI and AIA observations. Thus, each timestamp is represented as a variable-sized set of 1-24 AARPs, with an average of approximately 7 AR cutouts per timestamp. The distribution across our datasets can be seen in Figure 5.

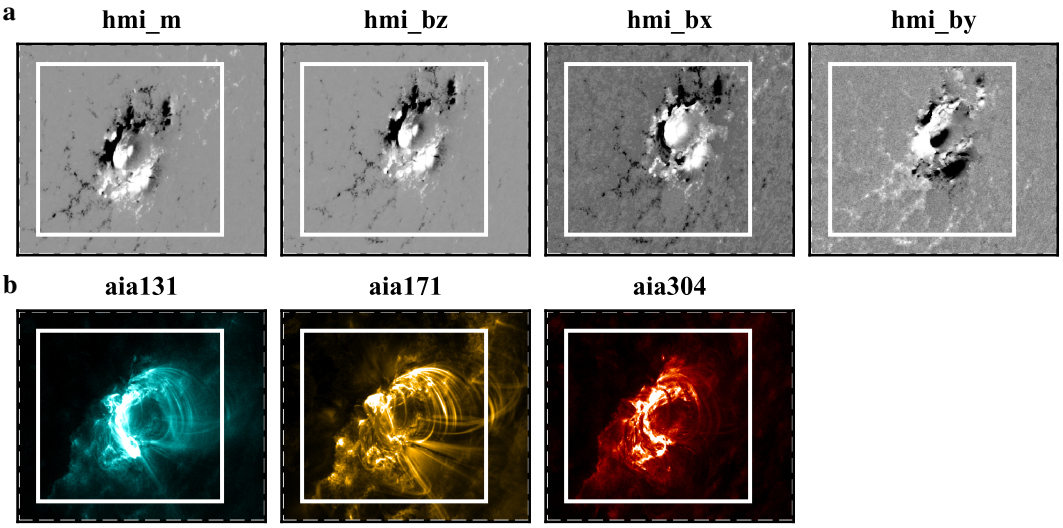


Figure 3: SHARP 7115 cutout image HMI and AIA channels at 2017/09/07 01:00

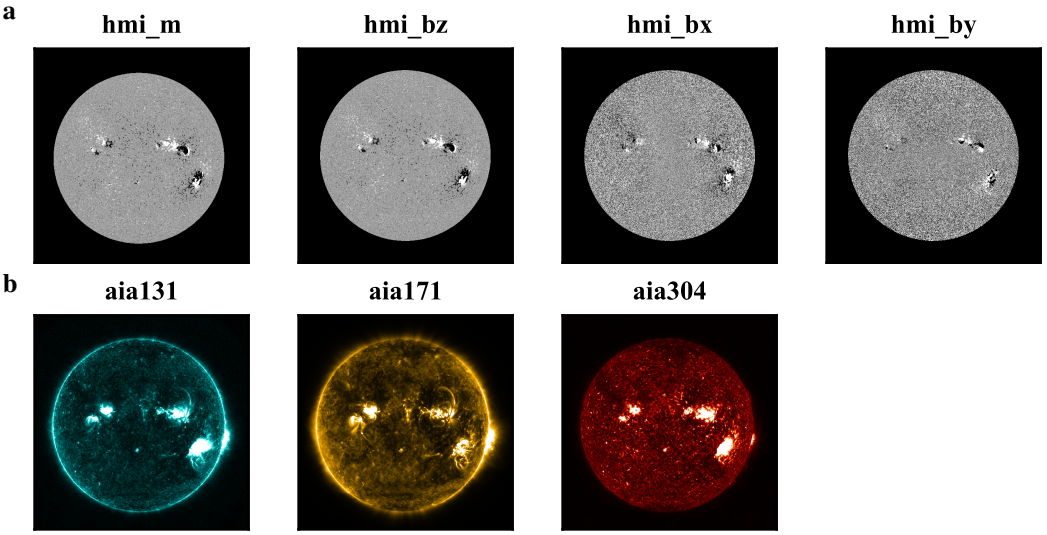


Figure 4: Solar full disk images of four HMI and three AIA channels at 2017/09/07 01:00. hmi_m refers to the line of sight magnetogram.

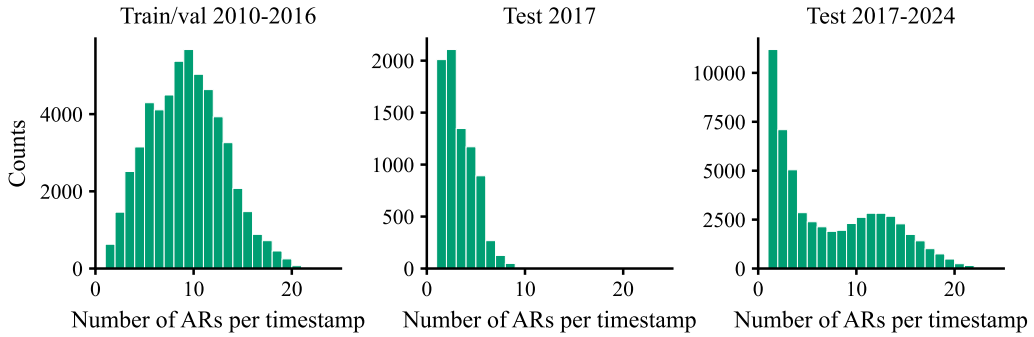


Figure 5: Number of ARs per timestamp across different datasets.

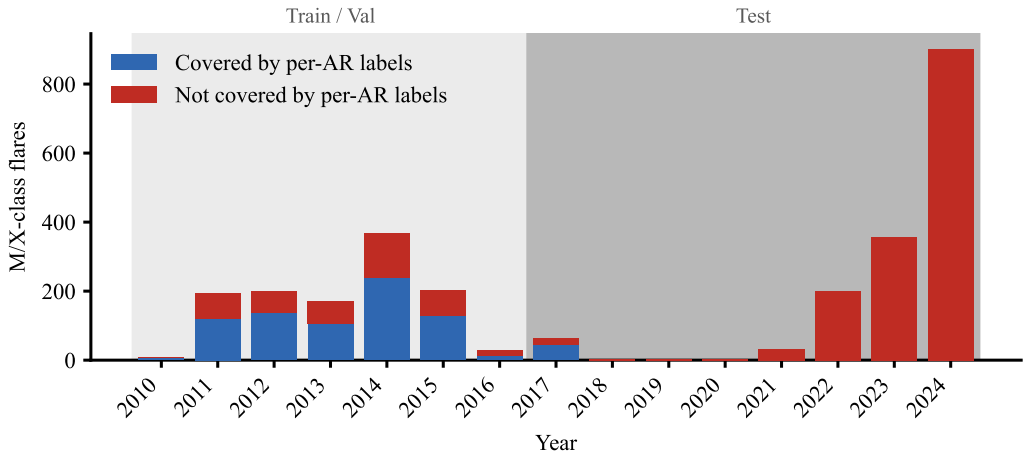


Figure 6: Number of flares per year and dataset split. We use the data from 2010-2016 for training and validation. The data from 2017-2024 is used as the holdout test-set. On a subset of the 2017 test data (blue) we evaluate the attention weights for each region as this overlaps with our flare catalog with AR labels [9].

Data normalization and spatial preprocessing. The different observational channels exhibit substantially different pixel-value distributions, and we therefore normalize each channel separately. The EUV intensities are strongly skewed towards higher values, with a distribution that is approximately exponential. We first apply a logarithmic transformation, $\log(1 + x)$, and subsequently standardize each channel to zero mean and unit variance. In contrast, HMI magnetograms contain both positive and negative values and exhibit a substantial noise-dominated regime around zero, approximately within ± 7 Gauss for the line-of-sight channel and ± 70 Gauss for the vector magnetograms. Applying a logarithmic transformation is therefore not appropriate, as it would disproportionately amplify the relative contribution of this noise-dominated regime. Instead, we clip the magnetogram values to the range $[-300, 300]$ Gauss and linearly rescale the resulting values to $[-1, 1]$. The clipping threshold is chosen such that the magnetic structures relevant for flare prediction remain well represented while limiting the influence of extreme values. The extracted AR patches have variable spatial dimensions. To obtain a fixed-size input compatible with the ResNet-18 architecture described in Appendix B, we resize all patches to 224×224 pixels. This resizing does not preserve the original spatial dimensions of the ARs, so we additionally investigated whether explicit information about the original patch position and size could improve the learned representation. Specifically, we encoded the original position and dimensions using a separate multilayer perceptron(MLP) and concatenated its output with the ResNet-18 embedding. This additional positional and scale information did not yield a significant improvement in forecasting performance and was therefore omitted from the final model.

Flare labels. Flares are categorized into A, B, C, M, or X-categories according to the peak magnitude of the soft X-ray flux. We restrict our analysis to M- and X-class flares (M+), corresponding to a peak flux of at least 10^{-5} W/m^2. Energetic events are particularly relevant for space-weather applications because their electromagnetic radiation can affect the Earth's ionosphere and disrupt radio communication and space-based technologies. The temporal distribution of flares over the period of our dataset can be seen in Figure 6.
We obtain flare start times and classifications from the NOAA GOES XRS flare catalog [18]. It is based on Observations of the Geostationary Operational Environmental Satellite system (GOES), specifically the X-Ray Sensor (XRS).
We label a selected observation timestamp as positive if at least one M/X-class flare occurs in subsequent 24 h, otherwise the bag is assigned a negative label. Therefore, the forecasting task is to predict wether a M+ flare will occur within the next 24 h solely based on the ARs present at the observation time. For training and validation we use the data from 05/2010–11/2016 and reserve the period from 2017–2024 as a held-out test set. Within the training and validation set we perform a five fold temporal cross-validation: For each fold, two calendar months from each year are assigned to the validation set (e.g. January and February). The two months immediately adjacent serve as temporal

buffer, while all the remaining months are used in the training set. For example, when January and February are used for validation, December of the preceding year and March are excluded and April to November are used for training. The temporal buffer reduces the risk of information leakage caused by the strong temporal correlation between neighboring observations and by the persistence of active regions across multiple days and weeks. The 2017–2024 period is not used for model selection and serves as a future test set, allowing us to evaluate wether the learned representation generalizes to observations from a later time period.
Because reliable associations between individual flares and their source ARs are available only for a limited subset of the data, we do not use AR-level flare-origin annotations to construct the training, validation, or test labels. Instead, all labels are derived exclusively from the global GOES flare catalog, consistent with the MIL formulation described in Section 2. The resulting class distributions differ substantially between the temporal splits; detailed class-balance statistics and additional evaluation results are provided in Appendix E.

Flare origin. To evaluate whether the learned instance weights identify the AR responsible for a predicted flare, we additionally use the SWAN space-weather dataset [9], which provides associations between individual flare events and their source active regions. The dataset is based on GOES XRS flare observations and combines information from multiple flare and AR catalogs to assign an AR origin to individual events. The resulting flare-AR associations are available from 2010 through 2017. We match the flare events in the SWAN catalog to the events in the NOAA GOES XRS catalog using a temporal tolerance of ± 1 hour. With this tolerance we cover approximately 65% of flares observed during this period from the GOES XRS catalog. For source region identification evaluation and the encoder comparison, we discard the timestamps that have a subsequent flare eruption according to [18] but no AR source annotation according to [9]. There are no flares originating from different ARs within 24h in the relevant time-frame according to the SWAN catalog. This means that each of our positively labeled timestamps for this evaluation set can clearly be attributed to one AR.
Importantly, these source-region annotations are used only for evaluation and are not used during training. Because the availability of reliable flare–AR associations is limited to this subset, we restrict the evaluation of the learned AR-level weights to the subset of M+ flares in 2017 for which a source-region annotation is available. This means the period of 2017 can be used for bag level flare prediction evaluation and also provides an independent set of flare-AR associations against which the attention weights can be evaluated. That way, we can assess whether the AR receiving the highest model weight corresponds to the independently annotated flare-producing AR, despite the model having been trained exclusively with bag-level flare labels.
The number of flares per year, in addition to and the availability of corresponding AR origin annotations, can be seen in Figure 6. Usually humans experts can visually inspect a series of images to decide which AR a flare is originating from. Such manual associations are labor-intensive and become increasingly challenging during periods of high solar activity when multiple ARs are simultaneously present. The SWAN catalog provides an automated alternative based on the cross-matching of multiple observational catalogs, but its incomplete coverage motivates restricting source-region evaluation to the annotated subset.

B Architecture details

Encoder network. We choose a standard ResNet-18 [16] as an encoder to for the AR image due to its versatility and low parameter count, as we only have access to limited data. We change the number of input channels in the first convolutional layer to accommodate all the seven channels from A. This ResNet-18 has 11.2 million parameters in our seven channel configuration.

Attention-based aggregation. As described in Section 2, ARENA represents each observation at time t as a bag of N_t active-region instances with corresponding embeddings $\mathbf{h}_i \in \mathbb{R}^M$. The attention-based multiple-instance learning (ABMIL) aggregator computes a scalar attention score for each instance and uses the normalized scores to form a weighted sum of the instance embeddings:

$$\mathbf{z} = \sum_{i=1}^{N_t} \alpha_i \mathbf{h}_i, \tag{4}$$

where the attention weights are already normalized via the softmax function in order to make the magnitude of the aggregation vector $\mathbf{z}$ independent of the number of instances inside a given bag.

For our main ARENA model, we use the gated attention mechanism proposed by Ilse et al. [12]. For each instance embedding $\mathbf{h}_i \in \mathbb{R}^M$, a non-normalized attention score for a give embedding $\mathbf{h_i}$ is computed as

$$a_\phi(\mathbf{h_i}) = \mathbf{w}^\top \left[\tanh\left(\mathbf{V}\mathbf{h}_i^\top\right) \odot \mathrm{sigm}\left(\mathbf{U}\mathbf{h}_i^\top\right)\right], \tag{5}$$

Here, $\mathbf{V}, \mathbf{U} \in \mathbb{R}^{L\times M}$ and $\mathbf{w} \in \mathbb{R}^L$ are learnable parameters, M is the dimension of the instance embedding, L is the dimension of the attention space, and $\odot$ denotes element-wise multiplication. The functions $\tanh(\cdot)$ and $\mathrm{sigm}(\cdot)$ denote the hyperbolic tangent and sigmoid nonlinearities, respectively. The gating mechanism combines two independently transformed versions of the instance embedding through element-wise multiplication before projecting the result to a scalar attention score. This allows the attention function to model more complex relations rather than relying on a single nonlinear transformation of the embedding. The resulting aggregation is invariant to the ordering of the active regions and naturally accommodates observations containing different numbers of ARs. The magnitude of the aggregated vector $\mathbf{z}$ is also independent of the sequence length due to the softmax normalization of the weights.

For our experiments, the ResNet-18 encoder produces embeddings of dimension $M = 512$ and we use an attention dimension of $L = 128$. The gated attention module therefore contains

$$2LM + L = 2 \cdot 128 \cdot 512 + 128 = 131{,}200, \tag{6}$$

trainable parameters, excluding any parameters associated with the ResNet-18 encoder and downstream classification head.

The aggregation vector $\mathbf{z}$ is then passed into a small classifier network which consists of two linear layers, projecting the 512-dimensional down to 256 dimension, applying the ReLU activation function and projecting into a single logit for classification. This small classifier network has $512 \cdot 256 + 256 \cdot 1 = 131{,}328$ parameters.

C Training details

Since M+ events constitute only a minor fraction of Sun observations, the forecasting dataset is strongly imbalanced. We therefore investigated several strategies for addressing class imbalance during training, including negative-instance undersampling, weighted random sampling, and class-weighted loss functions. Among these approaches, undersampling the negative training observations to obtain a balanced training set yields the best validation performance. Undersampling is applied only to the training data; the validation and test sets retain their original class distributions. Without undersampling, we observe a strong tendency toward the majority-class solution, with models assigning negative predictions to nearly all observations. Weighted random sampling and using a class-weighted loss reduce this effect but do not provide a consistent improvement over negative undersampling.

Therefore, we use negative undersampling for all experiments reported in this paper. We selected the optimization hyperparameters based on the validation performance during the development period. The final configuration uses a learning rate of $\eta = 10^{-7}$, a weight decay of 0.001, and a batch size of 32. Higher learning rates resulted in rapid degradation of validation performance, consistent with overfitting.

D Evaluation

Binary prediction metrics. We evaluate binary M+ flare forecasting using the True Skill Statistic (TSS), Heidke Skill Score (HSS), precision and F1-score. All three metrics are computed from the confusion matrix consisting of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Unless stated otherwise, the predicted class is obtained by applying the same decision threshold of 0.5 to the model output probability on the evaluation sets.

The TSS is defined as

$$TSS = TPR + TNR - 1 = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} - 1 \in [-1, 1], \tag{7}$$

where TPR and TNR denote the true-positive rate (recall) and true-negative rate (specificity), respectively. TSS ranges from -1 to 1, with 1 corresponding to perfect classification and 0 corresponding to no discriminatory skill under the standard interpretation. Unlike accuracy, TSS is insensitive to

changes in the class prevalence because it gives equal weight to the true-positive and true-negative rates. In theory, this makes it particularly suitable for evaluating flare forecasting under strong class imbalance [28].
The HSS measures the improvement in classification performance relative to the expected performance of random classification given the observed marginal class frequencies. We compute it as

$$HSS = \frac{2(TP \cdot TN - FP \cdot FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)} \in [-1, 1]. \tag{8}$$

An HSS of 1 indicates perfect classification and 0 indicates no improvement over the corresponding chance agreement. Unlike TSS, HSS depends on the class distribution and can therefore be affected by changes in event prevalence between datasets.
Additionally, we report the precision, defined as

$$Precision = \frac{TP}{TP + FP} \in [0, 1]. \tag{9}$$

Precision measures the fraction of predicted flares that correspond to actual flares. It therefore directly quantifies the reliability of positive predictions: low precision indicates a high rate of false alarms among predicted flare events.
The F1 score is the harmonic mean of the TPR and the precision:

$$F1 = 2 \cdot \frac{Precision \cdot TPR}{Precision + TPR} = \frac{2TP}{2TP + FP + FN} \in [0, 1]. \tag{10}$$

The F1 score also ranges from one, indicating perfect precision and recall, to zero if either precision or TPR is zero.

Instance-level encoder baseline. The instance-level classifier, whose performance is reported in Table 1 in Section 4, uses the same ResNet-18 encoder and classification MLP as ARENA to predict M-class flares from individual AR cutout images. Unlike ARENA, which learns from bag-level event labels through attention-based aggregation, the encoder baseline is trained using explicit AR–flare associations from the SWAN catalog [9]. Both models are trained and evaluated using the same temporal splits.
A direct comparison of their performance metrics must account for the different class prevalences in the two datasets: Although both datasets contain the same number of positive examples, these examples have different levels of granularity: positive AR instances in the instance-level dataset correspond to positive bags in the ARENA dataset. In the instance-level dataset, each AR cutout is labeled individually, whereas ARENA assigns a single label to the entire bag of AR cutouts. Consequently, AR instances in a positive bag that are not identified as flare-associated on an instance-level contribute as negative examples to the instance-level dataset, increasing the number of negative examples for the instance-level classifier. As a result, positive instance examples account for only 1.2 % of the instance-level dataset, compared with 3.86 % positive bags of the ARENA dataset. Consequently, the comparison should not be interpreted as a fully controlled benchmark as HSS, F1, and precision depend heavily on the class composition of the dataset. Nevertheless, it provides an indication of whether the weakly supervised aggregation introduces a substantial performance limitation with regard to the TSS, TPR, and TNR metrics.

Source region identification. As described in Section 3, we evaluate whether the attention-based ranking of active regions is consistent with the independently annotated flare source. For each positive observation, the ARs are ranked according to their normalized attention weights $\{\alpha_i\}$, and the resulting ranking is compared with the annotated source ARs recorded in the SWAN flare-AR catalog [9].
This evaluation is restricted to the temporally held-out 2017 test subset for which flare-AR associations are available, see Appendix A. For this specific subset of 2017, we discard all positive timestamps where the source region is unknown according to the SWAN catalog but keep all the negatives. The flare-AR annotations are not used during training, model selection, or attention weight computation. They are used exclusively for this post-hoc evaluation. Thus, source-region identification provides an independent assessment of whether the weakly supervised attention mechanism assigns high importance to the AR associated with the subsequent flare.
Identifying the flare origin is also a goal of the work by Francisco et al. [25] which introduced

Patch-CNN (P-CNN). There, the authors cut the full disk image into fixed rectangular tiles, apply a flare prediction network on the individual tiles and then use the maximum output probability over the different tiles to train their network on global flare prediction. The authors use EUV brightness center-of-mass calculations during a flare to identify its exact source region and compare it to post-hoc GradCam visualizations to evaluate their source region identification. However, the specific dataset and chosen evaluation period do not allow us to make a detailed comparison which will be explained in detail when the baseline approaches are introduced. Pandey et al. [29] used attention modules on the feature maps of different convolutional blocks. However, they only evaluate individual cases qualitatively without a global quantitative evaluation. Pandey et al. [21] and Yi et al. [30] used GradCam based methods to locate relevant features post-hoc after training the network for flare prediction without including a quantitative evaluation.

Flare prediction baselines. We compare ARENA results with three baselines: We (1) train a full-disk model from scratch, (2) fine-tune the Surya foundation model, and (3) report the results of P-CNN [25]. As a representative and directly comparable baseline, we train a full-disk model using the same seven observational channels as ARENA. We spatially downsample the 4096×4096 full-disk observations from the Surya dataset to 224×224 pixels and process them with the same ResNet-18 encoder [16] and classification MLP used by ARENA. This provides a direct comparison between the proposed representation as a variable-sized set of high-resolution AR cutouts and a single low-resolution full-disk representation while keeping the image encoder and classification head fixed. Additionally, we fine-tune the Surya foundation model [3] for M+ flare forecasting. Surya is pretrained using a pixel-level next-image prediction objective on multiple solar-observation channels, including the EUV channels used in our input representation. Because EUV observations can contain signatures of solar flares, evaluating on observations that were included in the Surya pretraining data would result in pretraining exposure to the evaluation period, resulting in data leakage. We therefore restrict evaluation to observations that were not used during Surya pretraining.
In the original pretraining procedure, days 16–30 of each year from 2010–2019 are used for validation, days 1–15 and 31–45 serve as temporal buffers, and the remaining observations before 2020 are used for pretraining. Observations from 2020–2024 constitute the held-out period. For our downstream flare-prediction experiment, we retain this temporal structure but restrict supervised training to observations from 2010–2016 and also undersample the negatives during training to match the training configuration for ARENA as closely as possible. We subsequently evaluate the fine-tuned model on observations from 2020–2024.
Finally, we compare against P-CNN [25] on flare prediction. P-CNN divides each full-disk solar observation into rectangular patches, processes each patch independently with a classifier, and uses the maximum patch-level probability as the global flare prediction. Since the authors do not provide an implementation, we use the performance values reported in their publication rather than re-training a model ourselves. They train separate models on three EUV channels and one the line-of-sight HMI magnetograms on data from 2010-2019 and report results on the test set from January 2020 through April 2023. We use the results of their best model (EUV), and compare our models on the corresponding period for comparison. They report their cross-validation results as the results of one ensemble model by averaging the probability outputs of their five models trained on different cross-validation folds. Therefore we do the same for our ARENA and full-disk model. For Surya this is not feasible due to aforementioned data leakage and limited computational resources.
The low-resolution full disk baseline is the most relevant to ARENA, as the results of Surya and Patch-CNN are not directly comparable to ARENA due to the different data splits and pre-processing procedures. For our Surya and low-resolution full-disk implementation, we only used the timestamps for which there are ARs according to [17] in order to replicate the dataset statistics. We compare the models on the same subset of 01/2020–04/2023 as the metrics change significantly across different datasets with different class balances which we will discuss later in Appendix E.

Additional evaluation analyses. In addition to the primary metrics reported in the main text we perform several analyses to characterize model performance under different evaluation conditions. We report confusion matrices and additional classification metrics for the validation folds and the different temporally held-out test subsets. We further investigate the relationship between class prevalence and the resulting performance metrics. Finally we look into our best validation fold (September/October, HSS=0.47) on the test set and analyze the most prevalent failure mode, false positives, in more detail.

E Additional results

To contextualize the results shown in Figure 2a, we first discuss the confusion matrices across the different validation folds on the test set and its subsets in Figure 7 and provide additional metrics in Table 2. Secondly, we look into the relationship between class balance and the resulting performance metrics in Figure 8.

Flare prediction and cross-validation. The results across the five cross-validation folds and the temporally held-out test subsets are summarized in Table 2. The corresponding confusion matrices for the test set and its subsets are shown in Figure 7. Across the evaluation subsets, ARENA maintains relatively high true-positive and true-negative rates, indicating that the model correctly identifies a substantial fraction of both flaring and non-flaring observations at the selected operating point. The ratio of TP to FP is reasonable for the 2017–2024 test set, but decreases for the 2017 tests sets. This behavior is expected for rare-event forecasting: even with good TPR and TNR, the precision (Equation 9) can be low due to a large class imbalance. Consequently, the model can achieve a high TSS (Equation 7) while simultaneously exhibiting relatively low precision on the naturally imbalanced test distribution.
Interestingly, when excluding the positive timestamps without AR source labels as identified by [9] from the 2017 test set (see Figure 7, going from the middle row to the bottom row), we observe a reduction of approximately 20% in true positives (top left matrix entries) and 70% in false negatives (top right matrix entries). This indicates that the model has problems in the classifying flares where there is AR label according to our catalog [9]. Potentially, these could be quiet Sun flares which do not originate from an AR, but further investigations with a bigger sample size are needed to draw firm conclusions.

Extra scalar parameters. We additionally evaluate whether incorporating the SHARP magnetic summary parameters [17] or the spatial position and size of each AR into the embedding helps in flare prediction. Neither modification results in a consistent improvement over the image-only ARENA model and all results are reported for the image-only ARENA model.

Table 2: Additional performance metrics on the different evaluation sets in the format $median^{Q75}_{Q25}$.

Dataset	TSS↑	HSS↑	F1↑	Precision↑	TPR↑	TNR↑
Validation	$0.48^{+0.05}_{-0.08}$	$0.42^{+0.00}_{-0.09}$	$0.54^{+0.05}_{-0.01}$	$0.46^{+0.05}_{-0.04}$	$0.70^{+0.04}_{-0.05}$	$0.77^{+0.02}_{-0.02}$
Test 2017–2024	$0.61^{+0.01}_{-0.01}$	$0.45^{+0.01}_{-0.01}$	$0.62^{+0.01}_{-0.01}$	$0.47^{+0.01}_{-0.01}$	$0.94^{+0.00}_{-0.00}$	$0.67^{+0.01}_{-0.01}$
Test 2017	$0.69^{+0.00}_{-0.00}$	$0.36^{+0.01}_{-0.03}$	$0.41^{+0.01}_{-0.03}$	$0.28^{+0.00}_{-0.03}$	$0.82^{+0.00}_{-0.01}$	$0.88^{+0.00}_{-0.02}$
AR subset 2017	$0.79^{+0.01}_{-0.02}$	$0.33^{+0.01}_{-0.04}$	$0.37^{+0.01}_{-0.03}$	$0.23^{+0.00}_{-0.03}$	$0.92^{+0.00}_{-0.00}$	$0.88^{+0.00}_{-0.02}$

Baseline comparison. The baseline performance on the relevant subset of our test set can be seen in Table 3. Our ARENA model performs similar to our full disk baseline across most metrics, but it does not reach the performance of P-CNN and Surya. Interestingly enough, ARENA has the highest TPR. ARENA closely matching the performance of the full disk network suggests that no information is lost when only focusing on the AR cutouts instead of the full disk images. Due to the different training splits P-CNN and Surya are not ideally controlled comparisons. Surya also has roughly 20x more parameters and therefore even fine-tuning with undersampling requires significantly more resources.

Metric dependence on class imbalance. The evaluation subsets differ substantially in their proportion of positive observations, which provides an important confounding factor when comparing performance across datasets. We therefore analyze the relationship between the positive-class prevalence and the reported classification metrics for the five validation folds, the full test set, and its subsets. The results for TSS and HSS are shown in Figure 8.
We observe a pronounced increase in TSS for the more strongly imbalanced evaluation subsets. In particular, the 2017 and 2017 AR-annotated subsets achieve high TSS values over 0.69, compared to significantly lower values for the more balanced validation and full test sets. This behavior is not expected from class prevalence alone, since TSS is defined in terms of TPR and TNR and is theoretically prevalence-independent. This could mean that the operating characteristics of our

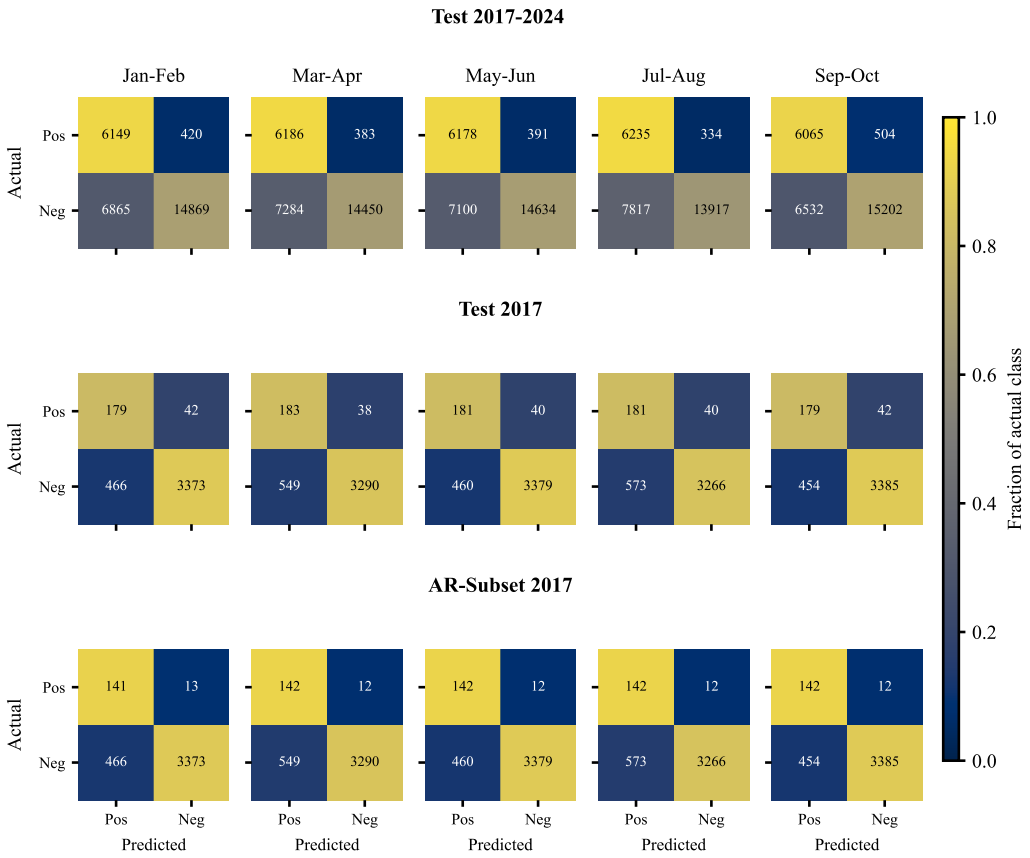


Figure 7: Confusion matrices for the different temporally held-out test subsets. The model correctly identifies a majority of both positive and negative observations, with consistent performances across each evaluation set. This can be observed when comparing TP and FN as well as TN and FP values, the model exhibits good predictive performance. This is also the case for the ratio of FN and TN values. However, comparing TPs and FPs, we observe that the model predicts a large number of FP values, especially in the 2017 test sets. The background colors are normalized per row of each matrix.

Table 3: Baseline comparison of ARENA on the January 2020–April 2023 subset of the test period, corresponding to the evaluation period reported for P-CNN [25]. ARENA and the full-disk ResNet, were implemented and trained by the authors; P-CNN results are taken from the original paper [25]. Pre-processing and training protocols differ between methods and therefore the comparison should not be interpreted as a fully controlled benchmark. Detailed experimental configurations are provided in Appendix D.

	Training set	TSS↑	HSS↑	F1↑	Precision↑	TPR↑	TNR↑
P-CNN (ensemble) [25]	2010–2019	**0.58**	0.42	0.53	0.39	0.82	0.76
Full disk (ensemble)	2010–2017	0.52	0.34	0.49	0.34	0.81	0.71
Fine-tuned Surya [3]	2010–2017	0.48	**0.46**	**0.55**	**0.54**	0.57	**0.90**
ARENA (ensemble)	2010–2017	0.51	0.30	0.46	0.31	**0.87**	0.64

model, particularly the true-positive and false-positive rates, vary across datasets with different class prevalences or that our model simply performs better on the 2017 evaluation sets. The high TSS observed for the 2017 AR subset should still be interpreted with caution, as it is estimated from a comparatively small number of positive observations.
For HSS, we observe a weaker tendency in the opposite direction: the highly imbalanced 2017 subsets achieve HSS values of approximately 0.29–0.37, whereas the full test set reaches approximately 0.45.

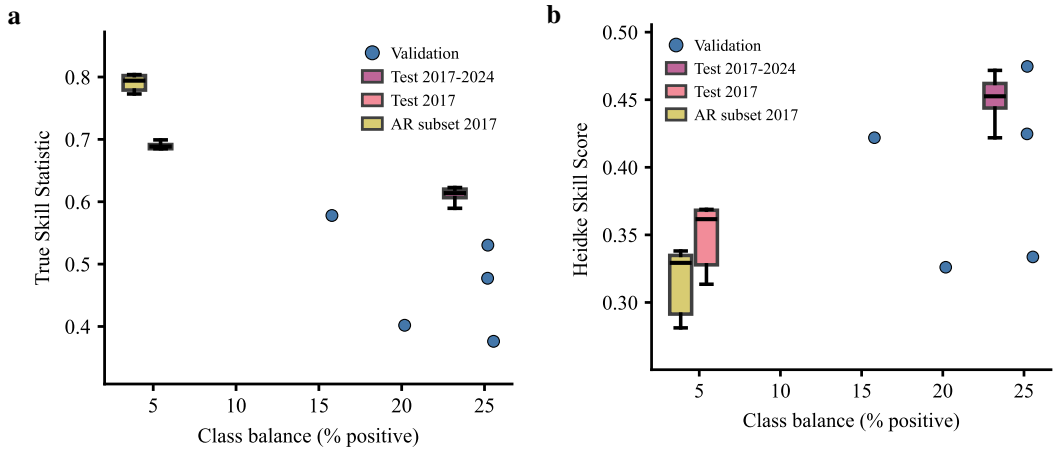


Figure 8: (a) TSS vs. positive-class prevalence: The evaluation subsets with lower positive-class prevalence tend to exhibit higher TSS, with the 2017 and 2017 AR-annotated subsets reaching values of approximately 0.69–0.80, compared with 0.40–0.62 for the other subsets. (b) HSS vs. positive-class prevalence: HSS shows a weaker tendency in the opposite direction, with values of approximately 0.29–0.42 for the highly imbalanced 2017 subsets and approximately 0.45 for the full test set. The validation subsets show no clear monotonic relationship.

The validation folds do not exhibit a clear monotonic relationship. These observations demonstrate that the choice of evaluation metric can substantially affect the apparent performance when the event prevalence differs between evaluation sets.

False-positive analysis. We further investigate the relatively large number of false positives observed for the 2017 test subset. As shown in the third row of Figure 7, this subset contains approximately three times as many false positives as true positives. The absolute number of false positives is partly a consequence of the strong class imbalance: even a comparatively small false-positive rate can result in many false-positive predictions when non-flaring observations greatly outnumber flaring observations. In addition, ARENA is trained using an approximately balanced sampling distribution, whereas the test set follows the substantially lower natural flare prevalence, resulting in a shift in class priors between training and evaluation.
To investigate whether the false positives exhibit additional temporal structure, we visualize the four classification outcomes over time for the 2017 test subset in Figure 9. It can be seen that false positives frequently occur immediately before or after sequences of true-positive observations. This suggests that a substantial fraction of the false positives are not isolated predictions but occur during periods in which the model assigns consistently high flare probability. To investigate the failure mode of temporally clustered false positives around the true positives further we visualize the temporal evolution of the output probability and the relevant attention weights of the cluster in early July visible in Figure 9. The resulting time-series is show in 10. We can clearly see one AR in yellow is dominating the attention. Timestamps leading up to the actual flares overlaid in orange are almost all correctly classified as positives. However we have significant false positives before and after the events. It appears that ARENA is learning an AR's elevated flare propensity over time, while the binary training target imposes a relatively sharp temporal boundary through the 24-hour prediction window.

F Related work

We already discussed Surya [3] and P-CNN [25] in the introduction of the baselines in section E. For image-based flare prediction, all approaches either use full disk images [2] or AR images with flare labels [5–7, 21, 22], as it is well-known that most flares originate from ARs [1]. Huwyler and Melchior [31] use MIL on different solar spectra to decide which frequency range is potentially informative for flare prediction.

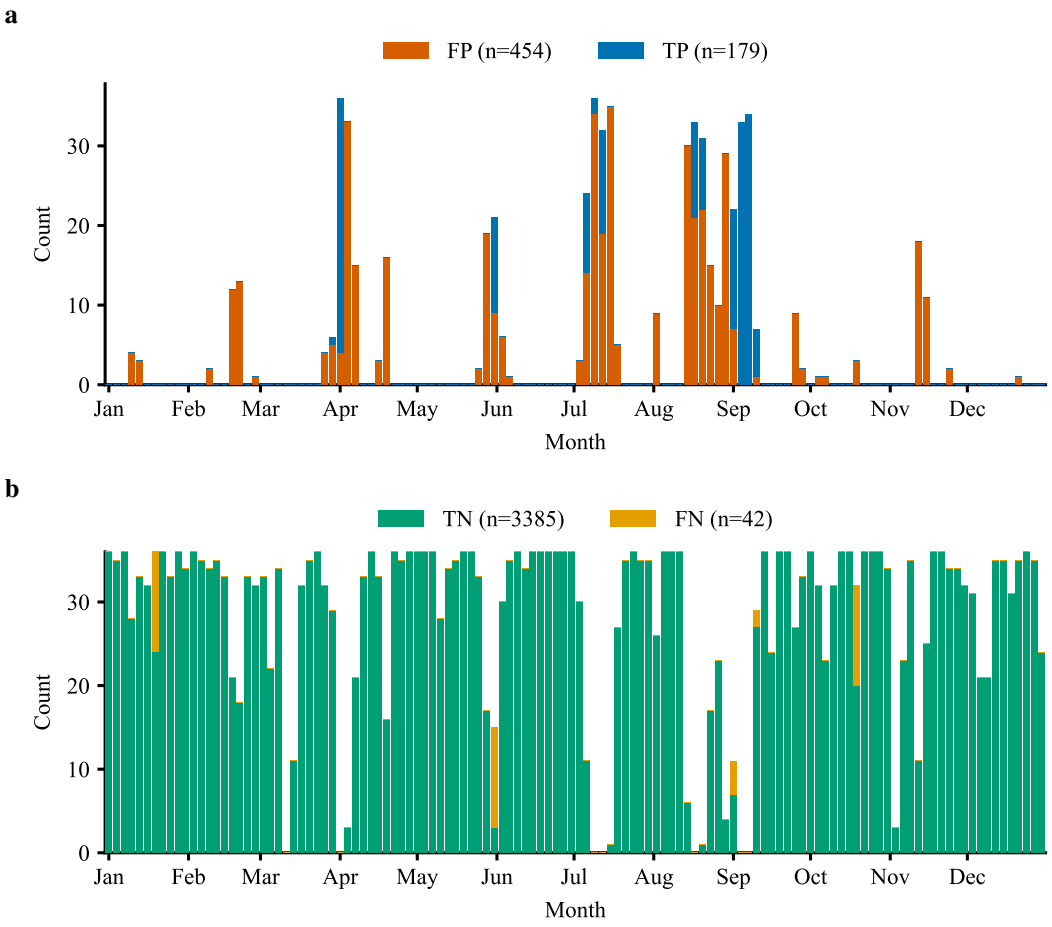


Figure 9: Temporal distribution of the four binary classification outcomes for the 2017 test subset. True positives and false positives are shown among the predicted-positive observations, while true negatives and false negatives constitute the predicted-negative observations. False positives frequently occur in temporal proximity to true positives.

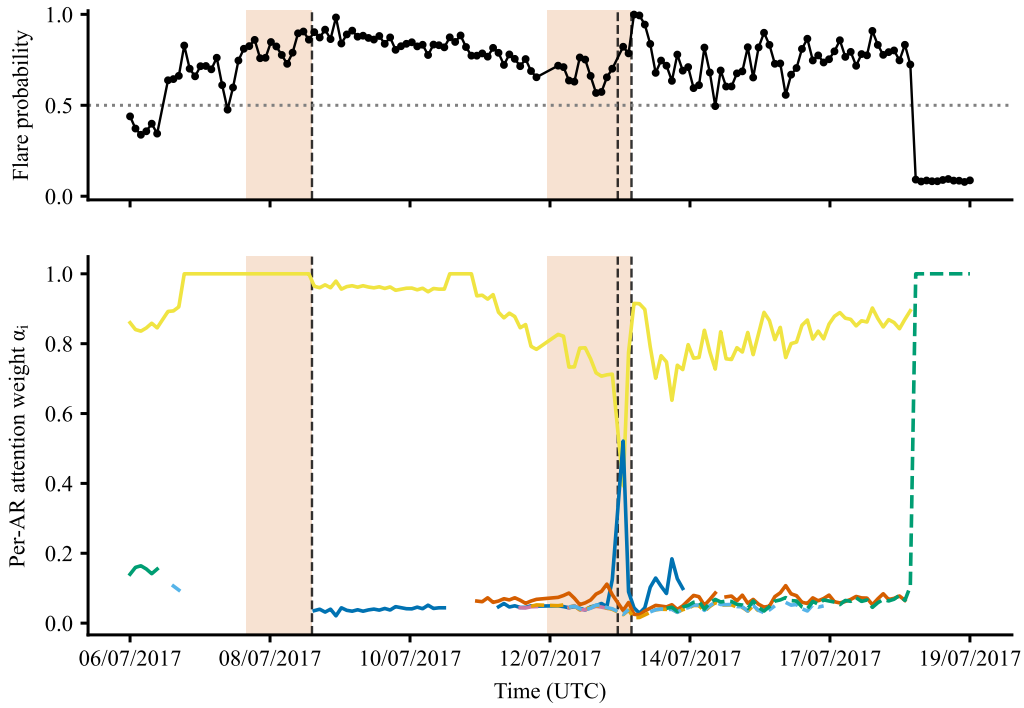


Figure 10: Temporal evolution of ARENA's predicted M+ flare probability (top) and attention weights assigned to individual active regions (bottom) for July 6th–19th, 2017. The orange shading indicates observations labeled positive according to our flare catalog because an M+ flare occurs within the subsequent 24-hour prediction window. Flares are represented by the black vertical dashed lines. The same active region receives high attention weights α_i across multiple consecutive observations surrounding the two flare events, while the predicted flare probability remains elevated over an extended period. As soon as this region disappears on July 18th, the output probability of the model drops. The yellow line corresponds to SHARP 7075 which is actually causing the two flares according to [9].