Cross-Modal Decoupled Knowledge Distillation for Imbalanced Solar Flare Prediction

Wantong Huang
Harbin Institute of Technology, Shenzhen
Shenzhen, Guangdong Province, China
2024311133@stu.hit.edu.cn

Yang Zhao
Harbin Institute of Technology, Shenzhen
Shenzhen, Guangdong Province, China
yang.zhao@hit.edu.cn

Yang Peng
The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong Province, China
224010018@link.cuhk.edu.cn

Abstract

Accurate and timely solar flare prediction is crucial for space weather forecasting. However, current operational systems face a critical performance tradeoff: high-dimensional image features offer superior accuracy but suffer from processing latency, while low-dimensional magnetic features are faster to process but lack high accuracy. To bypass these data-alignment bottlenecks, we propose cross-modal decoupled knowledge distillation (CMDKD) which employs a multimodal disentangled teacher to extract shared physical invariants during offline training, alongside asymmetric weighted optimization for extreme class imbalance. Evaluated on the Solar Cycle 24 dataset, CMDKD effectively injects high-dimensional spatial knowledge into an ultra-efficient 9D SHARP student model. By transferring the burden of multimodal integration to the offline training stage, our distilled single-modal network enables efficient inference without requiring strict multimodal synchronization, performing favorably against traditional baselines.

1. Introduction

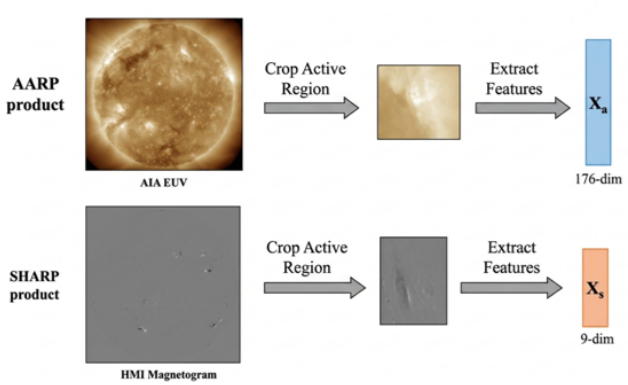
Solar flares are highly energetic explosive events that pose substantial risks to space-based assets, communication systems and broader space-based remote sensing [10, 14]. Accurate, timely prediction of these rare events is critical for disaster warnings, but operational forecasting faces a fundamental trade-off between predictive sensitivity and inference latency. High-dimensional multi-wavelength image statistics, such as the 176D AARP dataset (comprising 176 physically-interpretable parameters extracted from AIA active region patches) , offer rich spatiotemporal precursors but suffer from significant processing delays. Conversely, lightweight magnetic parameters such as a refined 9-dimensional SHARP subset of the space-weather HMI active region patches (SHARP) as illustrated in Figure 1, provide rapid operational responses but conventionally lack prediction accuracy.

To bridge this gap, multi-to-single cross-modal knowledge distillation (KD) [6] is promising but faces critical operational blind spots. First, direct cross-modal KD is often hindered by the modality gap [5] and the shortcut solution [3] problems. Heterogeneous noise misaligns soft labels, and joint training forces the teacher to prematurely shrink its capacity, degrading the guidance signal. Second, standard distillation objectives impose a conservative bias on rare flare events , prioritizing global accuracy at the cost of catastrophic missed flares [12].

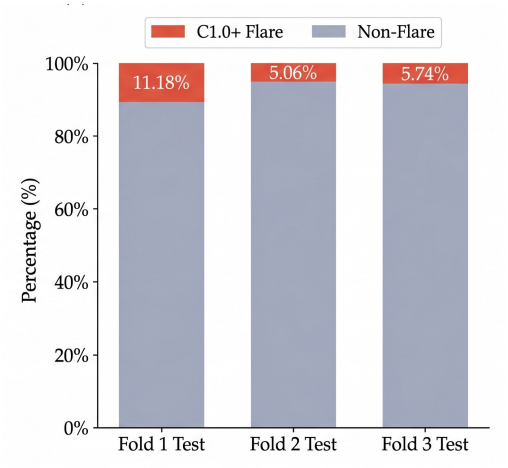
In this paper, we propose a cross-modal decoupled knowledge distillation (CMDKD) framework for imbalanced solar flare sequences as illustrated in Figure 1.We introduce a modality-disentangling teacher (MDT) that factorizes inputs into modality-specific and common latent spaces, isolating physical invariants from noise to transfer pure, stable soft labels. Combined with an asymmetric weighted optimization strategy for extreme class imbalance, these pure labels guide the lightweight 9-dimensional student to establish highly sensitive decision boundaries.

The main contributions are summarized as follows:

- We propose a novel multi-to-single knowledge distillation framework as illustrated in Figure 2 for solar flare prediction, successfully empowering a highly constrained 9-dimensional SHARP student with the high-dimensional spatial topology awareness of multi-wavelength imagery.
- We introduce a representation disentanglement mecha-



(a) Feature extraction process



(b) Class distribution

Figure 1. Overview of the multi-modal feature extraction pipeline and dataset distribution. (a) Illustration of the multi-modal active region feature extraction process based on AARP (AIA EUV) and SHARP (HMI Magnetogram) products. (b) Class distribution of C1.0+ flares versus non-flare samples across the three test folds.

nism to separate shared physical precursors, effectively filtering cross-modal noise and preventing teacher degradation during knowledge transfer.

- Our framework breaks the sensitivity-precision bottleneck under extreme class imbalance. Experimental results show that our distilled student achieves a state-of-the-art TSS of 0.7105(+3.7% relative) and an HSS of 0.3811(+32.7% relative) over the non-distilled model, significantly improving the robustness of operational forecasting.

2. Related Work

2.1. Solar Flare Prediction and Multimodal Integration

Recent data-driven solar flare prediction encompasses image-based learning and feature-based learning[9, 11, 13]. While deep learning on raw imagery captures fine spatial dynamics, feature-based learning remains favored operationally [4] to avoid the computational overhead of high-cadence streams. However, current models face a critical tradeoff. Computing comprehensive spatial statistics provide high predictive ceilings but introduce significant processing and cross-instrument synchronization delays. Crucially, traditional multimodal fusion demands strictly aligned multi-sensor data during inference—impractical given asynchronous satellite operations. Conversely, Near-Real-Time SHARP products offer a rapid pipeline but lack holistic spatial awareness. Bridging this gap to achieve multimodal performance using only a lightweight 9D SHARP student remains a challenge.

2.2. Learning under Extreme Class Imbalance

Solar flare prediction exhibits severe class imbalance, with positive events constituting less than 10% of the data [12]. Because standard objectives prioritize global accuracy at the cost of missed events, and direct optimization of non-smooth metrics e.g., TSS triggers gradient instability, we combine asymmetric weighted cross-entropy with distilled soft labels to establish a robust, smooth landscape for rare-event exploration [1].

2.3. Cross-Modal Knowledge Distillation and Disentanglement

Unlike traditional fusion that requires real-time data alignment, multi-to-single cross-modal KD distills knowledge from a high-dimensional teacher to a constrained single-modal student, completely bypassing synchronization bottlenecks and guaranteeing high-efficiency inference. However, direct cross-modal KD often suffers from the modality gap [5] and the shortcut solution problem [3], where joint training prematurely shrinks the teacher's representation capacity.

To solve these issues, representation disentanglement provides a structural solution. Recent frameworks, such as the incomplete cross-modal mutual knowledge distillation (IC-MKD) [11], utilize a modality-disentangling teacher (MDT) to separate modality-common from modality-specific features. Theoretical analyses confirm that cross-modal KD efficacy correlates positively with the transferred modality-common information. By isolating shared physical invariants, this disentanglement mechanism ensures the student receives pure, stable guidance without absorbing modality-specific noise.

3. Methodology

In this study, we propose a cross-modal decoupled knowledge distillation (CMDKD) framework for predicting solar flare class C1.0+ within a 24-hour horizon. Crucially, this

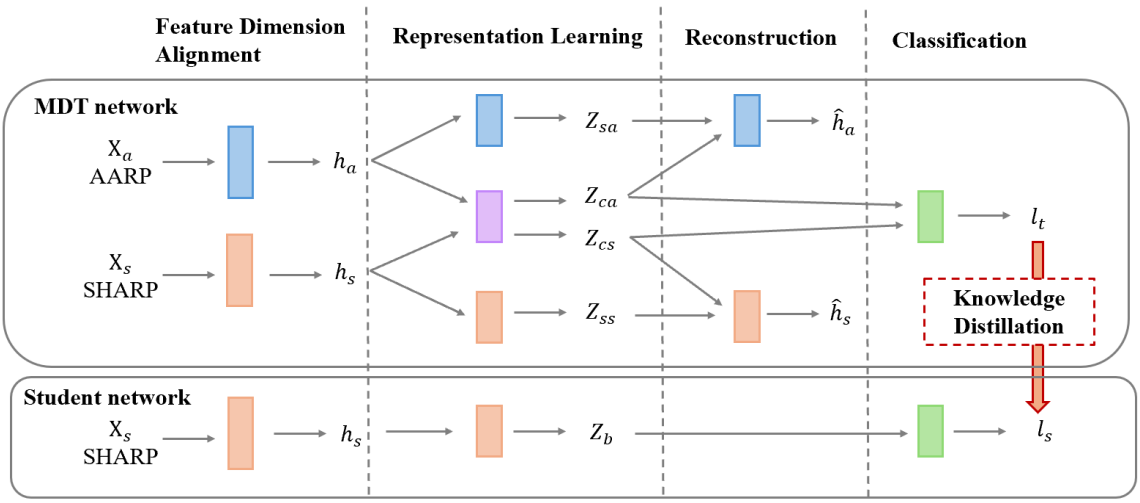


Figure 2. Teacher and student model architectures and training process in the CMDKD framework.

architecture is designed to bypass multi-sensor synchronization bottlenecks, enabling timely environmental monitoring via an ultra-efficient single-modal inference pipeline.

3.1. Decoupled Teacher Network

To leverage the complementary information of high-dimensional AARP and low-dimensional SHARP features without temporal alignment constraints, we construct the teacher model using a multimodal decoupled training (MDT) framework [7]. Initially, the 176-dimensional input $\mathbf{x}_a$ and 9-dimensional input $\mathbf{x}_s$ are mapped into a unified d-dimensional latent space ($d = 64$) via modality-specific MLP backbones, generating latent embeddings $\mathbf{h}_a$ and $\mathbf{h}_s$.

Decoupled Representation Learning. To isolate predictive flare precursors from modality-specific noise, each latent embedding $\mathbf{h}_m$ ($m \in \{a, s\}$) passes through two distinct encoding pathways. A modality-specific encoder captures the intrinsic physical properties unique to modality m, producing $\mathbf{z}_{sm}$ while a shared modality-common encoder extracts cross-modal precursor signals, yielding $\mathbf{z}_{cm}$.

Reconstruction & Classification. To minimize information loss and retain physical meaning, a decoder reconstructs the original latent vector $\hat{\mathbf{h}}_m$ by combining the common and specific representations ($\mathbf{z}_{cm} + \mathbf{z}_{sm}$). Conversely, the main classification task relies solely on the fused modality-common features ($\mathbf{z}_{ca} + \mathbf{z}_{cs}$). Isolating the classifier from modality-specific features ensures it learns only shared cross-modal knowledge, which is crucial for generating high-quality soft labels during knowledge distillation.

Overall Objective Function. The teacher network is optimized by a joint loss function combining four components to learn both decoupled features and robust classification boundaries:

$$\mathcal{L}_{MDT} = \mathcal{L}_{class} + \alpha_{sim}\mathcal{L}_{sim} + \alpha_{diff}\mathcal{L}_{diff} + \alpha_{recon}\mathcal{L}_{recon}. \tag{1}$$

The hyperparameter settings are detailed in Section 4. Specifically, the similarity loss ($\mathcal{L}_{sim}$) aligns common representations in a shared space by maximizing their cosine similarity to capture consistent cross-modal information:

$$\mathcal{L}_{sim} = 1 - \cos_{sim}(\mathbf{z}_{ca}, \mathbf{z}_{cs}). \tag{2}$$

The difference loss ($\mathcal{L}_{diff}$) applies soft orthogonality constraints to ensure independence between common and specific representations, forcing the encoders to focus on disjoint information:

$$\mathcal{L}_{diff} = \mathbf{z}_{ca} \cdot \mathbf{z}_{sa} + \mathbf{z}_{cs} \cdot \mathbf{z}_{ss} + \mathbf{z}_{sa} \cdot \mathbf{z}_{ss} + 3, \tag{3}$$

where the constant 3 ensures the loss remains non-negative. To avoid meaningless trivial features and ensure information completeness, the reconstruction loss ($\mathcal{L}_{recon}$) minimizes the mean squared error (MSE) between original and reconstructed latent vectors ($P = 64$):

$$\mathcal{L}_{recon} = \frac{1}{P}\sum_{i=1}^{P}(\mathbf{h}_{m,i} - \hat{\mathbf{h}}_{m,i})^2. \tag{4}$$

Finally, the classification loss ($\mathcal{L}_{class}$) applies a standard cross-entropy (CE) loss for the binary flare prediction. Crucially, to provide well-calibrated soft labels for the student, we do not apply extreme class re-weighting here, deliberately preserving the natural probability landscape of the highly imbalanced data.

3.2. Student Network

Architectural Design and Bottleneck Alignment. The backbone is an MLP ($d_{in} \to 128 \to 64 \to 32$), equipped with layer normalization and dropout to stabilize training under extreme class imbalance. Crucially, the 32-dimensional penultimate layer structurally aligns with the teacher's modality-common representation $\mathbf{z}_c \in \mathbb{R}^{32}$. This architectural symmetry creates a dedicated transfer channel, forcing the student to learn modality-invariant physical patterns rather than modality-specific noise.

Objective Functions for Asymmetric Learning. The student is trained with a composite loss that balances ground-truth supervision and cross-modal guidance:

$$\mathcal{L}_{\text{student}} = \mathcal{L}_{\text{WCE}} + \lambda \cdot \mathcal{D}_{KL}\left[\sigma\left(\frac{\mathbf{l}_T}{T}\right) \middle\| \sigma\left(\frac{\mathbf{l}_S}{T}\right)\right]. \quad (5)$$

The weighted cross-entropy ($\mathcal{L}_{WCE}$) applies an asymmetric weight to the positive class, heavily penalizing false negatives to establish a highly sensitive decision boundary for rare flares. Concurrently, the knowledge distillation ($\mathcal{D}_{KL}$) term transfers "dark knowledge" by softening the teacher and student logits ($\mathbf{l}_T$ and $\mathbf{l}_S$) with a temperature parameter $T > 1$. This smoothed soft-label guidance acts as a robust regularizer, bridging the informational gap between the lightweight SHARP input and the teacher's high-dimensional perspective. Ultimately, this enables the student to achieve pseudo-multimodal predictive performance using only a highly efficient single modality during inference.

4. Experiments and Results

4.1. Experimental Setup

Datasets and Chronological Splitting. We evaluate our framework using the high-dimensional AARP (176-dimensional) and low-dimensional SHARP (9-dimensional) datasets, derived from asynchronous satellite observations spanning Solar Cycle 24 (2010–2018)[2, 8]. To rigorously simulate real-world operational constraints and prevent data leakage, we adopt a strict chronological splitting strategy rather than random cross-validation. The dataset is partitioned into three distinct folds testing different phases of the solar cycle: Fold 1 (ascending to maximum), Fold 2 (maximum to declining), and Fold 3 (declining to minimum), as illustrated in Table 1.

Evaluation Metrics. Solar flare prediction is a severely imbalanced binary classification task, with positive events accounting for only 5%–15% of the data [12]. Under such extreme sparsity, standard accuracy (ACC) is heavily skewed toward the majority class. Therefore, we adopt the true skill statistic (TSS) as our primary metric to evaluate sensitivity to rare events. To comprehensively assess the model's ability to balance precision and recall, we supplement our analysis with the heidke skill score (HSS) and the F1-Score.

Table 1. Chronological dataset partitioning for the three cross-validation folds.

Fold	Train	Validation	Test
Fold 1	2010.1 – 2013.12	2014.1 – 2014.12	2015.1 – 2015.12
Fold 2	2010.1 – 2014.12	2015.1 – 2015.12	2016.1 – 2016.12
Fold 3	2010.1 – 2015.12	2016.1 – 2016.12	2017.1 – 2018.12

Implementation Details. For the teacher network, the balancing weights are set to $\alpha_{sim} = 10, \alpha_{diff} = 5, \alpha_{recon} = 10$. For the student network, we apply an asymmetric positive sample weight $\omega = 20.0$ in the WCE loss. The cross-modal knowledge distillation is governed by a temperature $T = 2.0$ and a balancing weight $\lambda = 0.3$.

4.2. Comparison with Traditional ML Baselines

To evaluate the predictive performance of the CMDKD framework, we compare our distilled SHARP student against three widely-used traditional machine learning baselines (SVM, Random Forest, and XGBoost), all trained on the identical 9-dimensional SHARP feature set.

In operational space weather forecasting, optimizing purely for global Accuracy or F1-Score under extreme class imbalance yields overly conservative models. For instance, in Table 2, Random Forest achieves the highest accuracy (94.25%) but exhibits a catastrophic TSS (0.3535) and recall (0.3665), missing most flares and rendering it impractical. While XGBoost provides a middle ground (TSS = 0.6616) and SVM demonstrates better sensitivity (TSS = 0.6999), these traditional baselines ultimately hit a performance bottleneck inherent to non-distilled single-modal learning.

Conversely, our proposed distilled SHARP model breaks this bottleneck to establish a superior operational equilibrium. It achieves a state-of-the-art TSS of 0.7105 and a robust recall of 0.8631, guaranteeing minimal missed detections. While maintaining competitive overall metrics (accuracy = 84.90%, F1-Score = 0.3811), our model firmly prioritizes the True Skill Statistic—the paramount metric in mission-critical environments where the catastrophic cost of a missed flare vastly outweighs a false alarm.

4.3. Effectiveness of Cross-Modal Distillation

To evaluate the specific contribution of cross-modal distillation in circumventing data alignment bottlenecks, we directly compare the distilled student networks against their baseline counterparts. As shown in Table 3, the "Base"

Table 2. Comparison of our proposed distilled SHARP student against traditional machine learning baselines. Best results are highlighted in **bold**.

Model	TSS ↑	HSS ↑	Recall ↑	F1-Score ↑	Accuracy ↑
SVM	0.6999	0.3932	0.8382	0.4537	86.06%
Random Forest	0.3535	**0.4427**	0.3665	0.4703	**94.25%**
XGBoost	0.6616	0.4264	0.7661	**0.4801**	88.64%
Ours (CMDKD)	**0.7105**	0.3811	**0.8631**	0.4440	84.90%

Table 3. Effectiveness of cross-modal distillation for both SHARP and AARP student networks and CMDKD demonstrates the student neworks which use teacher-guided KD loss.

Model	TSS ↑	HSS ↑	Recall ↑	F1-Score ↑	Accuracy ↑
SHARP (Base)	0.6854	0.2871	**0.9235**	0.3657	77.37%
SHARP (CMDKD)	**0.7105**	**0.3811**	0.8631	**0.4440**	**84.90%**
AARP (Base)	0.7364	0.3764	**0.8993**	0.4400	84.05%
AARP (CMDKD)	**0.7454**	**0.4404**	0.8654	**0.4949**	**87.81%**

models are trained solely with class-weighted cross-entropy to combat imbalance, whereas the "CMDKD" models incorporate the teacher-guided KD loss.

An analysis of the independent modalities reveals a critical tradeoff: relying exclusively on class weights forces naive over-prediction. For instance, the SHARP (Base) model achieves an exceptionally high recall (0.9235) but suffers from severe false alarms (HSS = 0.2871). Here, teacher-guided distillation (CMDKD) acts as a powerful structural regularizer. By accepting a marginal decrease in Recall (0.8631), the distilled SHARP model significantly curtails false positives, driving holistic improvements: TSS surges to 0.7105, Accuracy to 84.90%, and HSS to 0.3811. This robust regularization effect extends to the high-dimensional AARP modality. The distilled AARP student comprehensively outperforms its base counterpart, improving TSS from 0.7364 to 0.7454 and HSS from 0.3764 to 0.4404, while achieving a marked F1-Score improvement (0.4949) and an accuracy peak of 87.81%. .

Ultimately, Table 3 validates the core efficacy of CMDKD. Relying exclusively on weighted loss functions yields unbalanced models prone to false alarms. Cross-modal distillation resolves this by successfully transferring spatial and physical invariants from the multimodal teacher to the student. This establishes a vastly superior equilibrium between sensitivity and precision, entirely bypassing the need for synchronized multi-sensor inputs during deployment.

5. Conclusion

In this paper, we proposed the cross-modal decoupled knowledge distillation (CMDKD) framework to address modality disparity and extreme class imbalance in space-based remote sensing. By shifting the multimodal integration burden entirely to an offline training phase, CMDKD successfully distills high-dimensional topological insights from heterogeneous inputs into a lightweight, single-modal student network. Extensive evaluations demonstrated that our primary student model, relying solely on 9-dimensional SHARP parameters, achieved a state-of-the-art TSS of 0.7105 and an HSS of 0.3811. Furthermore, the bidirectional distillation gains observed in AARP single-modal students confirm the framework's universality. Crucially, by empowering accessible 9-dimensional magnetic parameters to rival high-dimensional imaging models, CMDKD provides a highly efficient and deployable solution for real-time space weather forecasting.

Limitations and Future Work. Currently, our teacher relies on compressed 176-dimensional statistical representations which restricts the exploitation of fine-grained spatial topologies and continuous temporal dynamics inherent in raw multi-wavelength satellite imagery. Future work will extend beyond this validation. We plan to integrate high-resolution raw image sequences and explore advanced cross-modal distillation frameworks. Leveraging complex spatiotemporal information, we aim to scale this multi-to-single distillation paradigm to large-scale real-time Earth observation and disaster response.

References

[1] Dara Bahri and Heinrich Jiang. Locally adaptive label smoothing for predictive churn. *arXiv preprint arXiv:2102.05140*, 2021. 2

[2] MG Bobra, JT Hoeksema, X Sun, and HMI Magnetic Field Team. Sharp: space-weather hmi active region patches. *SDO-3: solar dynamics and magnetism from the interior to the atmosphere*, page 17, 2011. 4

[3] Mengxi Chen, Linyu Xing, Yu Wang, and Ya Zhang. Enhanced multimodal representation learning with cross-modal kd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11766–11775, 2023. 1, 2

[4] Varad Deshmukh, Thomas Berger, James Meiss, and Elizabeth Bradley. Shape-based feature engineering for solar flare prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15293–15300, 2021. 2

[5] Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16006–16015, 2024. 1, 2

[6] Dino Ienco and Cassio Fraga Dantas. Discom-kd: Cross-modal knowledge distillation via disentanglement representation and adversarial learning. *arXiv preprint arXiv:2408.07080*, 2024. 1

[7] Min Gu Kwak, Lingchao Mao, Zhiyang Zheng, Yi Su, Fleming Lure, and Jing Li. A cross-modal mutual knowledge distillation framework for alzheimer's disease diagnosis: Ad-

dressing incomplete modalities. *IEEE Transactions on Automation Science and Engineering*, 2025. 3

[8] KD Leka, Karin Dissauer, Graham Barnes, and Eric L Wagner. Properties of flare-imminent versus flare-quiet active regions from the chromosphere through the corona. ii. nonparametric discriminant analysis results from the nwra classification infrastructure (nci). *The Astrophysical Journal*, 942(2):84, 2023. 4

[9] Shunya Nagashima and Komei Sugiura. Deep space weather model: long-range solar flare prediction from multiwavelength images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9396–9405, 2025. 2

[10] Keahi Pelkum Donahue and Fadil Inceoglu. Forecasting solar flares with a transformer network. *Frontiers in Astronomy and Space Sciences*, 10:1298609, 2024. 1

[11] Mingfu Shao, Hui Wang, Yuyang Li, Jiaben Lin, Jifeng Liu, Baolin Tan, Juan Guo, Yin Zhang, Jing Huang, Jiangtao Su, et al. Jw-flare: Accurate solar flare forecasting method based on multimodal large language models. *arXiv preprint arXiv:2511.08970*, 2025. 2

[12] Jie Wan, Jun-Feng Fu, Jin-Fu Liu, Jia-Kui Shi, Cheng-Gang Jin, and Huai-Peng Zhang. Class imbalance problem in short-term solar flare prediction. *Research in Astronomy and Astrophysics*, 21(9):237, 2021. 1, 2, 4

[13] Yun Yang, Yi Wei Ni, PF Chen, and Xue Shang Feng. Predicting solar flares using a convolutional neural network with extreme-ultraviolet images. *The Astrophysical Journal*, 985 (1):104, 2025. 2

[14] Wei Zhou, Yongqiang Yuan, Chengpan Tang, Yinan Meng, and Ying Chen. Ionosphere disturbances on gnss signal and positioning performance: analysis of the solar flare and geomagnetic storm events in september 2017 and october 2021. *Advances in Space Research*, 73(9):4608–4620, 2024. 1