Multi-Wavelength Transformer-Based 24-Hour Solar Flare Forecasting at the Active-Region Level

Dunia Alatoom [1], **Nikos Nikolaou** [2]

[1]Department of Artificial Intelligence, University of Jordan, Amman 11942, Jordan
[2]Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom

Key Points:

- Short sequences of recent multi-wavelength SDO observations contain sufficient information to forecast ≥M-class flares within 24 hours.
- Attention-based temporal modeling improves detection while controlling false alarms under extreme event imbalance.
- Compact spatiotemporal representations provide a strong and stable baseline for short-term space-weather forecasting.

Corresponding author: Nikos Nikolaou, n.nikolaou@ucl.ac.uk

Abstract

Solar flare forecasting remains challenging due to the complex spatiotemporal evolution of solar active regions and the severe class imbalance associated with high-impact events. In this work, we investigate a transformer-based framework for active-region–level solar flare forecasting using short sequences of multi-wavelength observations from the Solar Dynamics Observatory. The proposed approach integrates pretrained Vision Transformer representations with lightweight convolutional processing, explicit temporal differencing, and attention-based temporal aggregation to examine the role of compact temporal context in short-term flare prediction.Forecasting is formulated as a binary classification task targeting the occurrence of $\geq$M-class flares within a 24-hour prediction horizon. Evaluation is conducted on the SDOBenchmark dataset using active-region–level aggregation and skill-based metrics commonly adopted in space-weather forecasting. The results indicate that combining spatial representations with explicit short-term temporal modeling can yield stable forecasting skill under strong class imbalance. Across multiple random seeds, the selected configuration attains a mean True Skill Statistic (TSS) of 0.81 ± 0.04 and a Heidke Skill Score (HSS) of 0.73 ± 0.05, alongside high detection rates and controlled false-alarm behavior.

Plain Language Summary

Solar flares are sudden explosions on the Sun that can disrupt satellites, radio communication, navigation systems, and power grids on Earth. Accurately predicting when strong flares will occur remains difficult, because large flares are rare and active regions on the Sun evolve in complex ways over time. In this study, we examine whether short sequences of recent solar images contain enough information to predict strong solar flares within the next 24 hours. We use observations collected by a space-based solar observatory that monitors the Sun at multiple wavelengths, capturing activity in different layers of the solar atmosphere. Instead of relying on long historical records, our approach focuses on compact time periods covering only the most recent hours before a prediction is made. We apply a modern machine learning method that learns patterns from images and how those patterns change over time. The method combines information about the structure of solar active regions with their short-term evolution. We evaluate its performance using a publicly available benchmark dataset designed for solar flare forecasting. Our results show that this approach can reliably identify active regions that are likely to produce strong flares within 24 hours, while keeping false alarms at a manageable level.

1 Introduction

Solar flares are intense eruptions on the Sun that release vast amounts of energy, electromagnetic radiation, and high-velocity particles into space (Shibata & Magara (2011); Benz (2017)). These events are frequently associated with large-scale solar magnetic disturbances such as coronal mass ejections (CMEs) (Forbes et al. (2006)). They are among the most energetic eruptive phenomena in the solar system and pose a persistent threat to modern technological infrastructure. Their impacts on satellite operations, radio communications, navigation systems, and electrical power grids depend strongly on flare intensity and timing (Thomson et al. (2005); Hapgood (2011); Schrijver et al. (2014)).Recent solar activity illustrates these risks; for example, X-class flares have been reported to produce R3-level radio blackouts, as documented by the NOAA Space Weather Prediction Center (NOAA Space Weather Prediction Center (2024)).

In space-weather research, solar flares are commonly classified into five categories (A, B, C, M, and X) based on their peak soft X-ray flux measured by the

GOES satellites, with each class representing an order-of-magnitude increase in emitted energy relative to the previous one (Garcia (1994)). While low-energy A-, B-, and most C-class flares generally have negligible effects on Earth, M- and X-class flares are capable of inducing severe geomagnetic disturbances, making their reliable prediction a high-priority objective (NOAA Space Weather Prediction Center (2024)).

Solar flares predominantly originate from magnetically complex regions on the solar surface known as active regions (ARs). These regions form through the emergence and interaction of magnetic fields with opposite polarities, leading to the accumulation of magnetic free energy that may later be released explosively through magnetic reconnection processes (Priest & Forbes (2002); Forbes et al. (2006)). Active regions are commonly associated with sunspots and represent sites of strong and highly structured magnetic fields (Solanki (2003)). Line-of-sight magnetograms are routinely used to observe active regions by measuring the spatial distribution and strength of the photospheric magnetic field, providing crucial insight into magnetic configurations that precede flare activity (Schou et al. (2012)).

Early solar flare prediction methods typically constructed numerical representations of magnetograms by extracting a small set of physically motivated quantities, such as magnetic flux levels, field gradients, and properties of polarity inversion lines. These quantities were then used to train standard machine learning classifiers, including neural networks, support vector machines, and random forests (Colak & Qahwaji (2009); Ahmed et al. (2013); Bobra & Couvidat (2015); Boucheron et al. (2015); Nishizuka et al. (2017)). While these models achieved useful levels of predictive skill, their performance was inherently constrained by the limited expressiveness of manually designed features, which cannot fully describe the complex spatial organization of solar magnetic fields.

The rapid progress of deep learning shifted the field toward data-driven representation learning, particularly through the use of Convolutional Neural Networks (CNNs) for image-based flare prediction (Yang et al. (2025)). Several studies employed CNN architectures to analyze magnetogram images either cropped around active regions or spanning the full solar disk to forecast flare occurrence across different energy thresholds and time horizons (Huang et al. (2018); Park et al. (2018); Li et al. (2020); Deng et al. (2021); Abed et al. (2021)). These approaches demonstrated that spatial magnetic structures learned directly from images contain valuable information for flare forecasting.

However, solar flares are not instantaneous events but are preceded by magnetic evolution processes that unfold over several hours or days (Tlatov et al. (2018)). This temporal dependency motivated the use of sequence-based models, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) architectures, to model time-series representations of active-region evolution (Liu et al. (2019); Platts et al. (2022); Z. Sun et al. (2022)). Subsequent works combined spatial and temporal learning by integrating CNNs with recurrent models, enabling the joint modeling of magnetic structure and short-term temporal dynamics (Guastavino et al. (2023)). More recently, three-dimensional CNNs have been proposed to directly learn spatiotemporal correlations from magnetogram sequences (P. Sun et al. (2022)).

Despite these advances, many existing deep-learning approaches remain dominated by convolutional or recurrent architectures. In parallel, transformer-based models have emerged as a powerful alternative for vision and sequence modeling tasks, owing to their ability to capture long-range dependencies through self-attention mechanisms while enabling efficient parallel computation (Kaneda et al. (2022)). Initial studies have begun to explore transformer architectures for solar flare

forecasting, demonstrating their potential for modeling both spatial and temporal dependencies in solar observations (Grim & Gradvohl (2024)).

Nevertheless, the majority of prior work frames solar flare forecasting as a binary classification task over short prediction horizons, most commonly 24–48 hours. Furthermore, the inherent rarity of M- and X-class flares introduces significant class imbalance, complicating model training and evaluation. These challenges highlight the need for forecasting approaches that focus explicitly on high-impact events while effectively leveraging short-term temporal context.

In this study, we focus on the prediction of ≥M-class solar flares within a 24-hour forecast window. By leveraging modern representation-learning techniques and attention-based temporal modeling, our approach aims to assess the extent to which compact sequences of recent observations contain sufficient predictive information for short-term solar flare forecasting at the active-region level.

Rather than proposing a fundamentally new forecasting paradigm, this study systematically evaluates how recent advances in representation learning and attention-based temporal modeling can be combined within a unified framework for short-term solar flare forecasting. The emphasis is on clarifying the contribution of individual architectural components and on examining the predictive value of compact temporal sequences of recent multi-wavelength observations under a standardized benchmark setting.

2 Dataset

This study uses the SDOBenchmark dataset (Aerni & Bolzern (2026)), a publicly available benchmark for data-driven solar flare forecasting constructed from multi-wavelength observations acquired by the Solar Dynamics Observatory (SDO).

For each solar active region (AR), the dataset provides temporally ordered sequences of co-registered images spanning multiple observational channels. These include extreme ultraviolet (EUV) and ultraviolet (UV) images from the Atmospheric Imaging Assembly (AIA), as well as photospheric continuum intensity images and line-of-sight magnetograms from the Helioseismic and Magnetic Imager (HMI). Together, these modalities capture complementary physical processes across different layers of the solar atmosphere.

All images are originally provided at a spatial resolution of 256×256 pixels and are resized to 128×128 for computational efficiency. Each input sample consists of a short sequence of $T = 4$ observations selected from the most recent measurements preceding a reference time t_0. Although SDOBenchmark includes observations extending up to approximately 276 hours before t_0, only this compact short-term temporal context is used in the present study to focus on immediate pre-flare evolution.

Missing wavelength observations at a given timestamp are handled by zero-filling the corresponding channel to preserve a consistent input tensor shape. Such cases occur in fewer than 0.3% of samples across all wavelengths, indicating near-complete spectral and temporal coverage.

Each sequence is associated with a binary flare label derived from GOES soft X-ray peak flux measurements. A positive label is assigned if at least one solar flare of class M or higher ($\geq 10^{-5}\,\mathrm{W\,m^{-2}}$) occurs within a 24-hour forecasting window following t_0. The dataset configuration adopted in this study is summarized in Table 1.

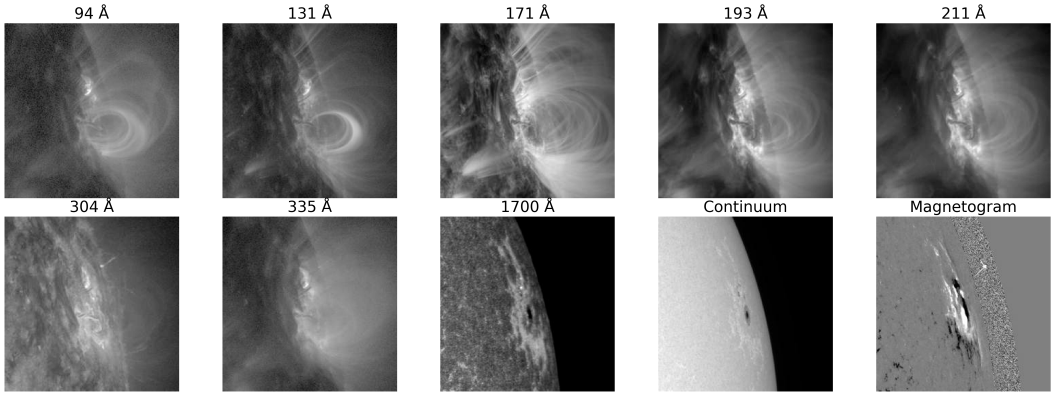

Figure 1: Example multi-wavelength observations of a solar active region from the SDOBenchmark dataset at a single timestamp.

Table 1: Summary of the SDOBenchmark dataset configuration used in this study.

Property	**Value**
Forecasting horizon	24 hours
Sequence length (T)	4 time steps
Input channels (total)	10
AIA EUV channels	94, 131, 171, 193, 211, 304, 335 Å
AIA UV channel	1700 Å
HMI channels	Continuum intensity, line-of-sight magnetogram
Image resolution	128×128 pixels
Flare labeling criterion	GOES peak flux $\geq 10^{-5}$ W m^{-2}

To prevent temporal or spatial information leakage, we adopt the predefined SDOBenchmark training and test partitions, which are defined at the active-region level. The training set is further subdivided to create a validation split, also enforced at the active-region level. The resulting dataset splits and class imbalance characteristics are reported in Table 2.

Table 2: Dataset splits and class distribution for the ≥M-class forecasting task.

Split	**Active Regions**	**Sequences**	**Positive Rate**
Training	491	3503	9.8%
Validation	164	1152	8.1%
Test	69	560	12.3%

The SDOBenchmark authors report two reference baselines on the public project repository: a fixed-point baseline (TSS = 0.0) and a first competitive model achieving TSS = 0.45 using a reduced set of input modalities. These baselines are intended to demonstrate dataset difficulty rather than to serve as fully optimized forecasting systems and are therefore not included in the comparative evaluation with peer-reviewed studies.

Multiple sequences extracted from the same active region may correspond to overlapping forecasting windows. To obtain active-region-level predictions, sequence-level probabilities are aggregated using a maximum rule. While SDOBenchmark provides a standardized and well-documented testbed, the results reported here should be interpreted as benchmark-specific performance estimates rather than evidence of universal generalization across instruments or solar cycles.

3 Method

We propose a hybrid deep-learning model for forecasting solar flares that learns from short sequences of multi-wavelength observations of solar active regions. The objective of the model is to capture how the spatial structure of active regions evolves over time and how this short-term evolution relates to the occurrence of major solar flares.

The design of the proposed model prioritizes architectural simplicity and transparency over complexity. Each component is selected to address a specific aspect of the forecasting problem: convolutional layers capture local spatial structure, pretrained Vision Transformers encode global spatial dependencies across the full active-region image, temporal differencing emphasizes short-term evolution, and the temporal transformer aggregates sequential information. The objective is not to introduce new architectural elements, but to systematically examine how these components interact when applied jointly to compact temporal sequences of solar observations.

Each input sample consists of a sequence of four observations collected prior to a reference time t_0. At each observation time, the active region is represented by ten co-registered solar images acquired at different wavelengths, covering coronal, chromospheric, and photospheric layers of the Sun. All images are resized to 128×128 pixels and processed on a per-image basis prior to temporal aggregation.

For each image, a lightweight convolutional stem is first applied to extract local spatial patterns and to reduce sensitivity to small-scale intensity variations. The

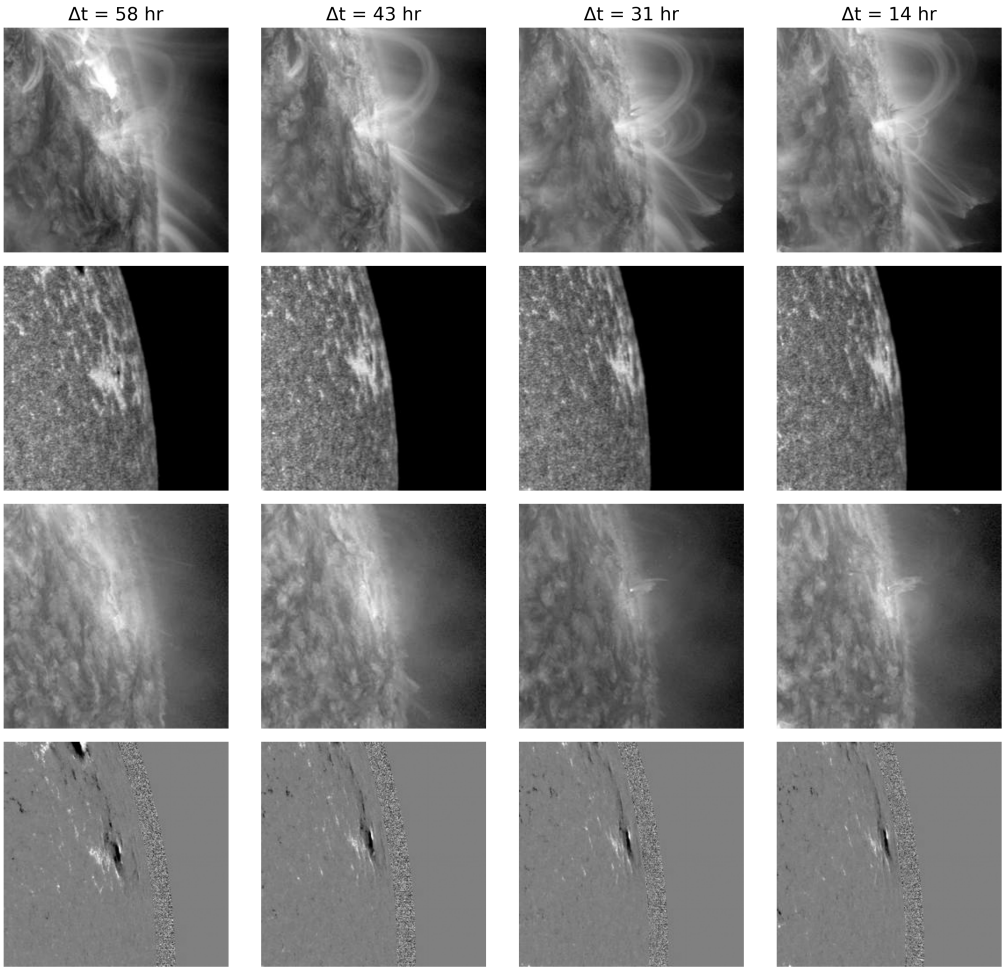

Figure 2: Example multi-wavelength temporal input sequence for a non-flaring active region, illustrating temporal evolution across spectral channels.

resulting features are then passed to a pretrained Vision Transformer (ViT) (Dosovitskiy et al. (2021)), which is used as a fixed spatial encoder. The ViT produces a compact global representation summarizing the spatial configuration of the active region at a given wavelength and time step.

At each time step, the representations from all wavelength channels are combined using mean pooling to form a single multi-spectral representation of the active region. This aggregation integrates complementary spectral information while maintaining a computationally efficient architecture.

To explicitly model short-term temporal changes, first-order temporal differences are computed between consecutive time steps. These temporal differences emphasize changes in the feature representations over time and are concatenated with the original features. A linear projection layer is then applied to map the combined features into a fixed-dimensional latent space suitable for temporal modeling.

The resulting sequence of feature vectors is processed by a transformer encoder operating along the temporal dimension (Vaswani et al. (2017)). Through self-attention, the temporal transformer learns dependencies across the four observation times, enabling the model to capture short-term temporal patterns that may precede solar flare activity.

Finally, a mask-aware pooling operation is applied across the time dimension to obtain a single sequence-level representation for each active region, where the pooling ignores padded or missing time steps arising from irregular temporal sampling. This representation is passed through a dropout layer for regularization and then through a linear prediction head to produce a scalar output. The output represents the predicted probability that the corresponding active region will produce at least one ≥M-class solar flare within the subsequent 24-hour forecast window. See Fig. 3.

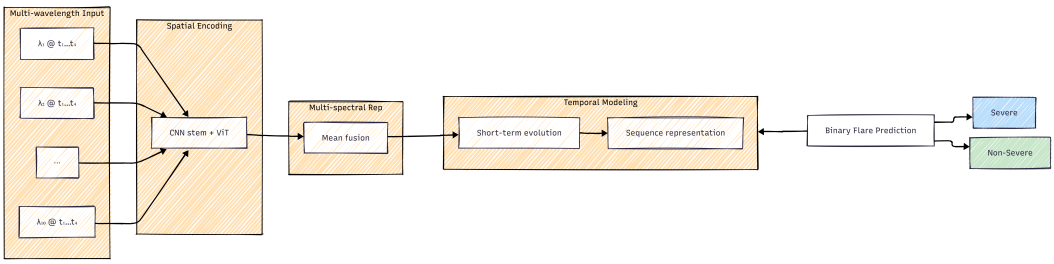

Figure 3: Overview of the proposed hybrid model architecture for active-region–level solar flare forecasting.

4 Experimental Setup

4.1 Data Splitting and Evaluation Protocol

To prevent information leakage between samples originating from the same solar active region, all data splits are performed at the active-region level rather than at the individual sequence level. Active regions are partitioned into training, validation, and test sets such that no region appears in more than one split. Stratified sampling is applied to preserve the proportion of flaring and non-flaring regions across splits.

Within each split, multiple temporal sequences may be extracted from a given active region. Model performance is evaluated at the active-region level by aggregating sequence-level predictions using a maximum operation, consistent with the forecasting formulation described in Section 2. This protocol reflects the operational objective of identifying flare-prone active regions rather than individual image timestamps.

4.2 Training Procedure and Hyperparameter Selection

Optimization is performed using the AdamW optimizer (Loshchilov & Hutter (2019)), which provides stable convergence for transformer-based architectures through decoupled weight decay. The pretrained Vision Transformer backbone is kept frozen and used as a fixed spatial feature extractor in order to reduce the number of trainable parameters and mitigate overfitting under severe class imbalance. A learning rate of 1×10^{-4} is adopted for all trainable layers, representing a conservative choice suitable for training task-specific heads on top of pretrained representations. Larger learning rates resulted in unstable validation behavior, while smaller values slowed convergence without improving forecasting skill. Training is conducted with a batch size of 32, selected as a compromise between gradient stability and computational efficiency.

A cosine learning-rate schedule with a 10% linear warm-up is applied to stabilize early optimization and promote smooth convergence(Loshchilov & Hutter (2017); Vaswani et al. (2017)). Weight decay is set to 5×10^{-4} to provide mild regularization. Mixed-precision training is employed to reduce memory usage and accelerate training. Early stopping is applied based on the validation-set True Skill Statistic (TSS), with a patience of five epochs. TSS is used instead of accuracy due to its robustness to class imbalance and its relevance for operational solar flare forecasting (Bloomfield et al. (2012)). The maximum number of training epochs is set to 30, although convergence is typically reached earlier.

Hyperparameter values are selected based on prior literature on transformer optimization and data-driven solar flare forecasting (Loshchilov & Hutter (2019); Bloomfield et al. (2012)), task-specific considerations such as severe class imbalance and limited positive samples, and empirical validation across multiple random seeds, with emphasis placed on training stability rather than exhaustive hyperparameter tuning.

4.3 Class Imbalance Handling and Loss Function

Solar flare forecasting is characterized by severe class imbalance, with flaring events occurring far less frequently than non-flaring samples. To mitigate the resulting bias during training, we adopt a combination of loss-based reweighting and sampling-based strategies.

We employ focal loss (Lin et al. (2017)), which dynamically down weights easy to classify samples and emphasizes harder examples that are more informative for learning. Focal loss is defined as

$$\mathcal{L}_{\text{focal}} = -\alpha(1 - p_t)^{\gamma} \log(p_t), \tag{1}$$

where

$$p_t = y\,p + (1 - y)(1 - p). \tag{2}$$

Here, $p \in (0, 1)$ denotes the predicted probability for the positive (flaring) class, $y \in \{0, 1\}$ is the ground-truth label, α is a weighting factor, and γ controls the degree of focusing on hard-to-classify samples. In all experiments, we set $\gamma = 2.0$ and $\alpha = 0.5$, following the standard focal loss formulation commonly used for

class-imbalanced classification (Lin et al. (2017)). We verified that moderate variations around these values did not lead to consistent improvements in validation performance.

In addition to loss-based reweighting, class-balanced sampling is applied during the initial training epochs using a weighted random sampling strategy, where sampling probabilities are inversely proportional to class frequencies. This warm-up phase increases exposure to minority-class sequences during early optimization and stabilizes gradient updates. After this phase, uniform random sampling is employed to preserve the natural data distribution and improve generalization.

Model outputs are continuous probabilities representing the likelihood of $\geq$ M-class flare occurrence within the 24-hour forecasting horizon. To convert these probabilities into binary predictions, a decision threshold is selected using the validation set.

Specifically, sequence-level probabilities are first aggregated to the active-region level using a maximum rule. The decision threshold is then chosen to maximize the True Skill Statistic (TSS) over the validation set. This threshold selection procedure is performed independently of the test set to prevent information leakage. The selected threshold is fixed and subsequently applied during test-time evaluation.

4.4 Threshold Selection and Evaluation Metrics

Rather than adopting a fixed classification threshold, the decision threshold is selected based on validation performance. Predicted probabilities on the validation set are aggregated at the active-region level and evaluated across candidate thresholds, with the threshold maximizing the validation True Skill Statistic (TSS) selected and fixed for test-set evaluation.

Model performance is evaluated using standard solar flare forecasting metrics, including TSS, HSS, POD, FAR, CSI, F1-score, accuracy, AUROC, and AUPRC. All metrics are computed at the active-region level following probability aggregation. Formal definitions are provided in Appendix Appendix A.

Table 3: Training and optimization hyperparameters used in this study.

Parameter	**Value**
Sequence length (T)	4
Number of wavelengths	10
Channels per image	3
Batch size	32
Image resolution	128×128
Maximum epochs	30
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	5×10^{-4}
Loss function	Focal loss ($\alpha = 0.5$, $\gamma = 2.0$)
Learning-rate schedule	Cosine decay with 10% warm-up
Early stopping	Validation TSS (patience: 5 epochs)

The threshold optimization procedure is intended to identify a representative operating point under the chosen evaluation protocol, rather than to claim

an optimal or universal decision threshold. Threshold-dependent metrics should therefore be interpreted as performance at a favorable validation-selected operating point, while probabilistic metrics such as AUROC and AUPRC provide a more threshold-independent assessment of model behavior.

5 Results

This section presents the results of the proposed method for solar flare forecasting under a 24-hour prediction horizon. All results are reported at the active-region level following the evaluation protocol described in Section 4. To account for stochastic variation during training, reported results correspond to the mean and standard deviation.

Table 4 summarizes the active-region-level performance of the proposed model across a range of threshold-based and probabilistic evaluation metrics. The model demonstrates strong and stable forecasting skill, achieving a mean True Skill Statistic (TSS) of 0.81 $\pm$ 0.04 and a Heidke Skill Score (HSS) of 0.73 $\pm$ 0.05, indicating substantial improvement over random and climatological baselines.

The model attains a high Probability of Detection (POD) of 0.93 ± 0.06, reflecting effective identification of flare-prone active regions, while maintaining a relatively low False Positive Rate (FPR) of 0.12 $\pm$ 0.05. This balance between detection and false alarms is further reflected in the Critical Success Index (CSI) and F1-score, highlighting robust discrimination between flaring and non-flaring regions at an operationally relevant decision threshold.

In addition, the model achieves strong probabilistic performance, with an AUROC of 0.96 ± 0.01 and an AUPRC of 0.89 ± 0.03, indicating reliable ranking of flare likelihoods across active regions despite significant class imbalance.

Table 4: Active-region-level forecasting performance for $\geq$ M-class solar flares within a 24-hour forecast horizon. Reported values correspond to the mean and standard deviation over six seeds.

Metric	**Value**
TSS	0.806 ± 0.041
HSS	0.731 ± 0.055
POD	0.931 ± 0.058
FPR	0.125 ± 0.047
FAR	0.283 ± 0.072
CSI	0.677 ± 0.052
F1-score	0.807 ± 0.036
Accuracy	0.889 ± 0.027
AUROC	0.959 ± 0.014
AUPRC	0.894 ± 0.030

Table 5 summarizes the performance of recent solar flare forecasting studies that rely on image-based, parameter-based, or hybrid input representations. Direct comparison across methods should be interpreted with caution, as the reported metrics are influenced by factors such as class imbalance, dataset selection, the choice and number of input channels (i.e., wavelengths), the inclusion of SHARP magnetic parameters, and the underlying data representation (images versus derived

Table 5: Comparison with representative solar flare forecasting studies for ≥ M-class flares within a 24-hour forecast horizon.

Paper	Data†	TSS	HSS	POD	FAR	CSI	F1
Nishizuka et al. (2018)	HMI-par	0.80	0.26	0.95	0.82	0.18	–
Huang et al. (2018)	HMI-img	0.66	0.14	0.85	0.90	0.10	0.18
Liu et al. (2019)	HMI-par	0.79	0.32	0.89	–	–	–
Li et al. (2020)	HMI-par	0.75	0.76	0.82	0.11	–	–
Wang et al. (2020)	HMI-par	0.68	0.38	0.73	0.05	–	–
Guastavino et al. (2022)	HMI-par	0.68	–	–	–	–	–
Kaneda et al. (2022)	HMI-par	0.53	–	–	–	–	–
Grim & Gradvohl (2024)	HMI-img+SHARP	0.70	0.25	0.79	0.83	0.16	0.27
Yang et al. (2025)	AIA(3)*	0.88	0.58	0.89	–	–	–
Roy et al. (2025)	AIA+HMI(13)	0.44	0.52	–	–	–	0.56
Our work	**AIA+HMI(10)**	**0.81**	**0.73**	**0.93**	**0.28**	**0.68**	**0.81**

Dashes (–) indicate metrics not reported in the original publications.
†HMI-img: HMI line-of-sight magnetogram images;
HMI-par: HMI SHARP or magnetic parameters;
AIA(N): AIA EUV images with N channels;
+SHARP: inclusion of SHARP parameters.
* Full-disk images

parameters) adopted in each study. In addition, not all studies report variability estimates across multiple training runs; several works report only single or optimal performance values, whereas the results presented in this work are reported as mean ± standard deviation over multiple random seeds.

Earlier approaches based on handcrafted magnetic parameters extracted from HMI magnetograms (e.g.,Nishizuka et al. (2018);Liu et al. (2019);Li et al. (2020)) generally achieve competitive detection rates, as reflected by high POD values in several cases. However, these models rely on predefined feature sets, which may limit their ability to capture the full spatial complexity of magnetic structures preceding flare activity. In contrast, image-based methods using raw magnetograms or EUV observations (e.g.,Huang et al. (2018); Yang et al. (2025)) aim to learn spatial representations directly from the data, often improving generalization to unseen active regions at the cost of increased false-alarm rates.

Hybrid approaches that combine image information with magnetic parameters, such as (Grim & Gradvohl (2024)), demonstrate that integrating complementary representations can lead to balanced performance across multiple metrics, including TSS and CSI. Similarly, multi-channel image-based models using EUV observations (e.g., Yang et al. (2025);Roy et al. (2025)) highlight the potential benefit of incorporating coronal information beyond photospheric magnetograms, although reported skill scores vary depending on the selected channels and model architectures.

Within this context, the results of the present work indicate a consistent balance between detection capability and false-alarm control. The reported TSS and HSS values suggest effective discrimination between flaring and non-flaring cases, while the relatively high POD indicates sensitivity to flare-producing active regions. At the same time, the FAR remains within a comparable range to other image-based

and hybrid methods, suggesting that improved recall is not achieved at the expense of excessive false alarms. These results reflect the contribution of combining multi-channel AIA and HMI observations with a unified spatiotemporal representation, rather than reliance on a single data modality or handcrafted parameters.

Overall, the comparison demonstrates that recent progress in solar flare forecasting arises primarily from improved data representations and modeling strategies, rather than a single dominant approach. The performance of the proposed method aligns with this trend, offering a competitive balance of forecasting skill while remaining consistent with the limitations and variability observed across existing studies.

Beyond threshold-dependent metrics, probabilistic forecast quality is evaluated using receiver operating characteristic (ROC) and precision–recall (PR) analyses. Figure 4 presents the mean active-region–level ROC and PR curves averaged over six seeds, with shaded regions indicating one standard deviation. The ROC curve demonstrates strong discriminative capability across a wide range of operating points, with consistently high true positive rates achieved at relatively low false positive rates. This behavior is reflected in a high mean area under the ROC curve (AUROC) of 0.959 ± 0.012, indicating robust ranking of flare-prone and non-flaring active regions and reduced sensitivity to the choice of decision threshold.

Complementary evaluation using the precision–recall curve highlights the model's effectiveness under severe class imbalance. The mean area under the PR curve (AUPRC) of 0.894 ± 0.030 indicates that the predicted probabilities remain informative even when focusing on high-confidence flare predictions. In particular, the PR curve maintains relatively high precision across a broad range of recall values, demonstrating the model's ability to identify rare $\geq$ M-class flaring regions while limiting false alarms.

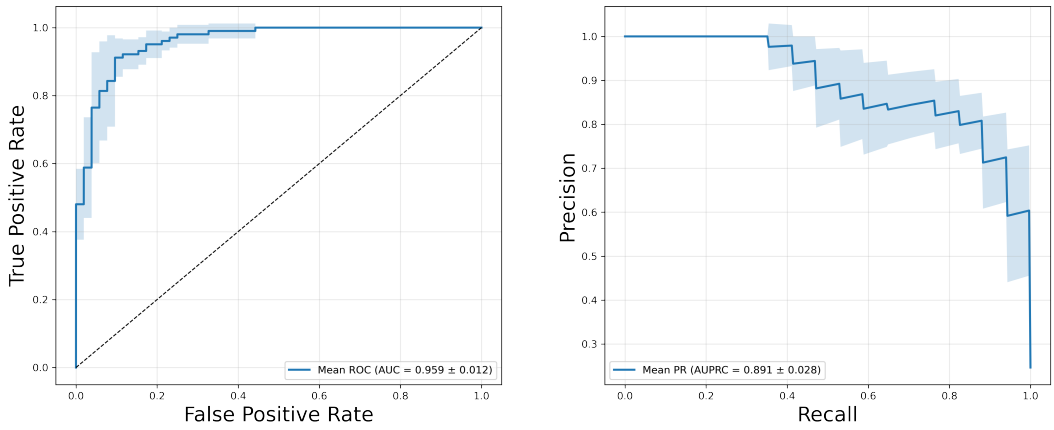

Figure 4: Mean AR-level probabilistic performance over six seeds for a 24-hour forecast horizon. Left: ROC curve. Right: precision–recall curve. Shaded regions indicate one standard deviation across random seeds.

Together, these results confirm that the proposed method produces discriminative probabilistic forecasts that are well suited for operational solar flare forecasting scenarios, where reliable ranking and confidence estimation are critical.

5.1 Active-Region-Level Aggregation

Aggregating predictions at the active-region level leads to more stable and interpretable forecasts compared to sequence- or frame-level evaluation. By considering the maximum predicted probability across multiple temporal sequences for each active region, the proposed approach effectively captures flare-prone behavior while reducing sensitivity to isolated noisy observations.

This aggregation strategy aligns model evaluation with practical space weather forecasting objectives, where the primary goal is to assess whether an active region poses an elevated flare risk within the forecasting horizon.

To further analyze the operational behavior of the proposed model, we examine the active-region–level confusion matrix on the test set for a 24-hour forecasting horizon. Since all quantitative results are reported as mean ± standard deviation across all training runs, confusion matrices are likewise averaged across seeds to reflect expected model behavior rather than a single stochastic realization.

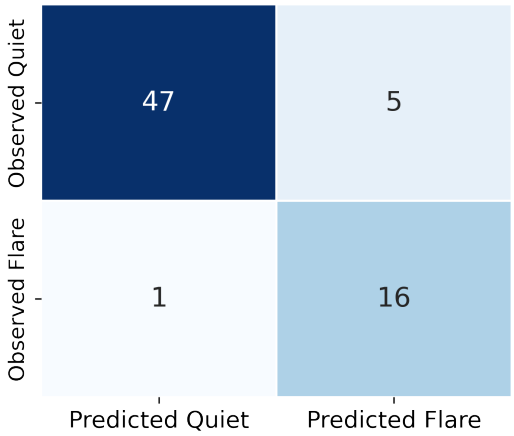

Figure 5: Mean active-region–level confusion matrix averaged across six seeds for the 24-hour forecasting horizon.

Figure 5 presents the mean active-region–level confusion matrix. The model correctly identifies the majority of flare-producing active regions, achieving a high true positive count while maintaining a relatively low number of false alarms. In particular, false negatives are rare, indicating that the model seldom fails to flag active regions that subsequently produce ≥M-class flares within the forecast window.

The number of false positives remains moderate, reflecting a controlled trade-off between sensitivity and specificity that is consistent with the reported values of TSS and FAR. Fractional values arise naturally from averaging confusion matrices across multiple runs and represent the expected number of cases per seed, rather than discrete outcomes from a single model instance.

To provide qualitative insight into the model's decision behavior, we further visualize active-region–level predicted flare probabilities aggregated across six seeds. Figure 6 shows the aggregated probabilities for each active region, ranked by ensemble confidence. All active regions are displayed as background points, while false positives and false negatives at the selected operating threshold are explicitly highlighted.

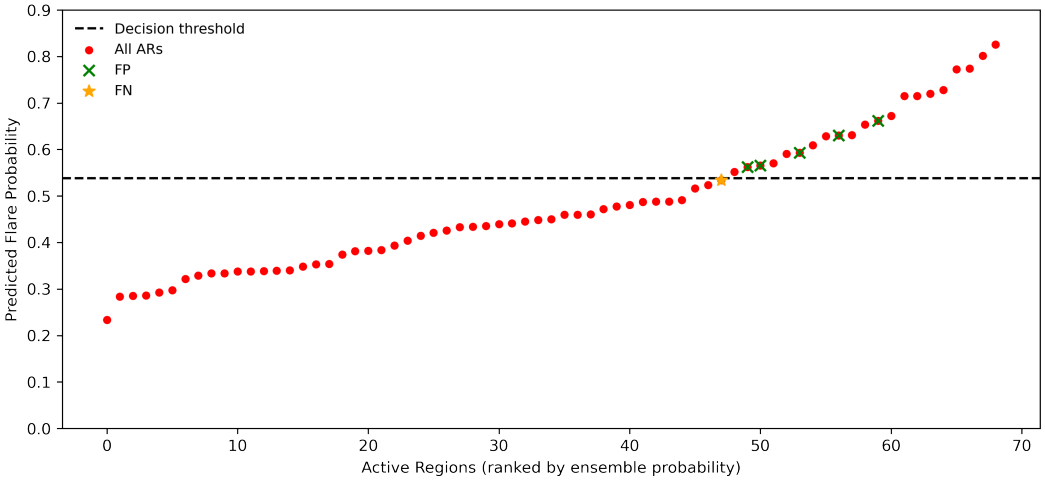

Figure 6: Active-region–level predicted flare probabilities for the 24-hour forecast horizon, aggregated across six seeds and ranked by ensemble confidence. False positives (FP) and false negatives (FN) are highlighted relative to the decision threshold selected on the validation set.

The figure reveals a clear separation between most flare-producing and quiet active regions, with the majority of false positives occurring close to the decision boundary. These cases correspond to active regions whose predicted probabilities lie near the selected threshold, reflecting intrinsic uncertainty rather than systematic misclassification.

From a physical perspective, such uncertainty is consistent with active regions that reside near the transition between flare-productive and non-productive states, where small scale or rapidly evolving magnetic changes, potentially unresolved at the available temporal cadence, can influence flare triggering. Additional sources of uncertainty may arise from projection effects and limitations in line-of-sight magnetic field measurements. In simple terms, most near-threshold cases correspond to active regions whose flare productivity is marginal, meaning that small or short-lived changes can determine whether a flare occurs within the forecast window.

False negatives are rare at the ensemble level, indicating that active regions missed in individual training runs are often correctly identified after aggregation across random seeds. This behavior suggests that ensemble aggregation stabilizes probability estimates and reduces sensitivity to borderline cases. In practice, such near-threshold predictions may be treated probabilistically or monitored with adaptive decision thresholds in operational settings, rather than interpreted as definitive binary outcomes.

6 Ablation Study

To assess the contribution of individual architectural components, we conduct an ablation study by incrementally extending a Vision Transformer (ViT) baseline. All model variants are evaluated on the 24-hour solar flare forecasting task using identical data splits, training procedures, and evaluation metrics. Performance is reported as the mean and standard deviation over three runs with different random seeds. Emphasis is placed on skill-based metrics commonly adopted in space-weather

forecasting, particularly the True Skill Statistic (TSS) and the Heidke Skill Score (HSS), which are robust to class imbalance.

We begin with a ViT-only baseline and progressively introduce a convolutional feature extractor (CNN), temporal intensity deltas, and a temporal transformer module. An additional variant incorporating wavelength attention is evaluated to examine whether explicit spectral reweighting provides further benefit.

Table 6: Ablation study results for the 24-hour solar flare forecasting task. All metrics are reported as mean ± standard deviation over three random seeds.

Metric	**ViT (baseline)**	**+CNN**	**+Δ**	**+Temp. Transf.**	**+WL Attn.**
TSS	0.578 ± 0.010	0.735 ± 0.092	0.756 ± 0.071	**0.781 ± 0.048**	0.722 ± 0.061
HSS	0.546 ± 0.005	0.660 ± 0.071	0.627 ± 0.110	**0.701 ± 0.071**	0.642 ± 0.014
POD	0.725 ± 0.028	0.882 ± 0.083	**0.980 ± 0.028**	0.922 ± 0.028	0.882 ± 0.096
FPR	0.147 ± 0.018	0.147 ± 0.009	0.224 ± 0.095	**0.141 ± 0.050**	0.160 ± 0.047
Precision	0.618 ± 0.019	0.660 ± 0.033	0.607 ± 0.099	**0.690 ± 0.085**	0.651 ± 0.051
FAR	0.382 ± 0.019	0.340 ± 0.033	0.393 ± 0.099	**0.310 ± 0.085**	0.349 ± 0.051
CSI	0.500 ± 0.000	0.609 ± 0.070	0.596 ± 0.087	**0.650 ± 0.063**	0.592 ± 0.015
Accuracy	0.821 ± 0.007	0.860 ± 0.028	0.826 ± 0.066	**0.874 ± 0.036**	0.850 ± 0.014
F1-score	0.667 ± 0.000	0.755 ± 0.052	0.743 ± 0.084	**0.786 ± 0.058**	0.743 ± 0.010
AUROC	0.925 ± 0.006	0.951 ± 0.012	0.957 ± 0.017	0.948 ± 0.009	**0.952 ± 0.003**
AUPRC	0.822 ± 0.010	0.878 ± 0.019	**0.890 ± 0.041**	0.876 ± 0.030	0.874 ± 0.010

The ViT-only baseline captures global spatial context but lacks explicit temporal modeling, resulting in limited forecasting skill and moderate recall. Introducing a CNN improves local spatial feature representation and leads to consistent gains across skill-based metrics, indicating that convolutional inductive biases complement transformer-based global attention.

Incorporating temporal intensity deltas substantially increases recall, with some runs achieving near-perfect detection of flaring regions. However, this gain is accompanied by increased false positive rates and higher variability across seeds, suggesting that delta features alone encourage more aggressive classification behavior.

The addition of a temporal transformer mitigates these effects by explicitly modeling sequential dependencies across delta-encoded representations. This configuration achieves the most balanced performance across evaluated metrics, improving overall skill while maintaining stable false-alarm rates.

Across all configurations, probabilistic ranking performance remains relatively stable, with incremental improvements observed when temporal information is introduced. The addition of wavelength attention marginally improves AUROC but does not translate into consistent gains in skill-based metrics.

Overall, the temporal transformer reduces inter-seed variability compared to delta-only configurations, indicating improved training stability. Based on these observations, we adopt the ViT+CNN+Δ+T configuration as the reference model for subsequent experiments, as it provides a balanced integration of spatial and temporal information without introducing unnecessary complexity.

7 Discussion

This work demonstrates that short-term solar flare forecasting benefits from explicit temporal modeling of recent multi-wavelength observations. By combining pretrained Vision Transformer–based spatial representations with lightweight convolutional features, temporal intensity differences, and a transformer-based temporal encoder, the proposed model captures both the structural state of solar active regions and their short-term evolution preceding flare onset.

Strong active-region–level performance indicates that compact temporal sequences spanning recent hours contain sufficient predictive information for identifying ≥M-class flares. High True Skill Statistic and Probability of Detection values show that the model learns discriminative temporal patterns associated with flare productivity while maintaining controlled false-alarm rates, a critical requirement under severe class imbalance.

The ablation study further clarifies the contribution of individual components. While spatial representations alone provide useful contextual information, temporal differencing substantially improves detection by emphasizing short-term evolution. Incorporating a temporal transformer stabilizes this behavior, reducing false alarms and yielding a more balanced and robust classifier. These results suggest that flare precursors are more effectively characterized by structured temporal evolution than by isolated snapshots.

Overall, attention-based temporal modeling combined with multi-wavelength solar observations provides a simple and effective framework for short-term active-region solar flare forecasting. Meaningful forecasting skill is achieved using compact temporal context and modest architectural complexity, making this approach a strong baseline for future data-driven space-weather studies.

8 Conclusion

This study presented a transformer-based framework for short-term solar flare forecasting at the active-region level using multi-wavelength observations from the Solar Dynamics Observatory. By integrating pretrained Vision Transformer representations with lightweight convolutional processing, explicit temporal differencing, and attention-based temporal aggregation, the proposed approach captures both spatial structure and short-term temporal evolution of solar active regions.

Experiments on the SDOBenchmark dataset (Aerni & Bolzern (2026)) demonstrate reliable forecasting skill for ≥M-class flares within a 24-hour prediction horizon. Across multiple random seeds, the selected configuration achieves a mean True Skill Statistic (TSS) of approximately 0.81 and a Heidke Skill Score (HSS) of approximately 0.73, indicating effective discrimination under severe class imbalance. High Probability of Detection is achieved while maintaining competitive false-alarm rates, together with strong probabilistic performance as measured by AUROC and AUPRC.

The ablation study confirms the importance of combining convolutional spatial encoding with explicit temporal differencing and transformer-based temporal modeling. While additional architectural extensions were explored, the selected configuration provides a favorable balance between model complexity and forecasting performance, supporting the use of compact temporal context rather than long historical sequences.

In comparison with prior image-based and hybrid approaches at similar prediction horizons (Nishizuka et al. (2018); Grim & Gradvohl (2024); Yang et al. (2025)),

the forecasting skill achieved here is competitive and, in several cases, superior. Unlike methods based on handcrafted magnetic parameters (Liu et al. (2019); Li et al. (2020)), the proposed framework learns unified spatiotemporal representations directly from multi-wavelength observations while avoiding the high false-alarm rates reported in some image-based models (Huang et al. (2018)). These findings align with recent work showing that advances in solar flare forecasting are driven primarily by improved data representations and temporal modeling strategies (Grim & Gradvohl (2024); Roy et al. (2025)).

Future work will extend this framework to full-disk solar observations to enable evaluation under more realistic operational conditions and facilitate direct comparison with recent full-disk forecasting studies (Yang et al. (2025)).

Acknowledgments

The authors would like to thank the developers of the SDOBenchmark dataset for providing a well-documented, standardized, and computationally accessible benchmark for solar flare forecasting research. This work made use of observations from the Solar Dynamics Observatory (SDO), including data from the Atmospheric Imaging Assembly (AIA) and the Helioseismic and Magnetic Imager (HMI).

9 Data and Code Availability

The SDOBenchmark solar flare forecasting dataset used in this study is publicly available at `https://i4ds.github.io/SDOBenchmark/` and has been cited in the References.

The code used for data preprocessing, model training, evaluation, and analysis is archived on Zenodo (Version v1.0.0) and is available at `https://doi.org/10.5281/zenodo.18600513` (Alatoom & Nikolaou (2026)).

10 Declarations

The authors declare there are no conflicts of interest for this manuscript.

11 AI tools disclosure

ChatGPT was used for language editing (grammar and readability) only. All scientific content, analysis, and conclusions were produced by the authors, who reviewed and approved the final manuscript.

Appendix A Evaluation Metrics

All evaluation metrics are computed at the active-region level following probability aggregation, as described in Section 4. Let TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively.

$$\mathrm{POD} = \frac{TP}{TP + FN}, \qquad \mathrm{FPR} = \frac{FP}{FP + TN}, \tag{A1}$$

$$\mathrm{FAR} = \frac{FP}{TP + FP}, \qquad \mathrm{CSI} = \frac{TP}{TP + FP + FN}, \tag{A2}$$

$$\mathrm{TSS} = \mathrm{POD} - \mathrm{FPR}, \tag{A3}$$

$$\mathrm{HSS} = \frac{2(TP \cdot TN - FP \cdot FN)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}. \tag{A4}$$

Probabilistic performance is assessed using the Area Under the Receiver Operating Characteristic Curve (AUROC) and the Area Under the Precision–Recall Curve (AUPRC).

References

Abed, A. K., Qahwaji, R., & Abed, A. (2021). The automated prediction of solar flares from sdo images using deep learning. *Advances in Space Research*, *67*(8), 2544–2557.

Aerni, M., & Bolzern, R. (2026). *Sdobenchmark: Solar flare prediction image dataset.* `http://i4ds.github.io/SDOBenchmark/`.

Ahmed, O. W., Qahwaji, R., Colak, T., Higgins, P. A., Gallagher, P. T., & Bloomfield, D. S. (2013). Solar flare prediction using advanced feature extraction, machine learning, and feature selection. *Solar Physics*, *283*(1), 157–175.

Alatoom, D., & Nikolaou, N. (2026). *solar-flare-forecasting-vit: v1.0.0.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.18600513` doi: 10.5281/zenodo.18600513

Benz, A. O. (2017). Flare observations. *Living reviews in solar physics*, *14*(1), 2.

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

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

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

Colak, T., & Qahwaji, R. (2009). Automated solar activity prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space weather*, *7*(6).

Deng, Z., Wang, F., Deng, H., Tan, L., Deng, L., & Feng, S. (2021). Fine-grained solar flare forecasting based on the hybrid convolutional neural networks. *The Astrophysical Journal*, *922*(2), 232.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International conference on learning representations (iclr)*. Retrieved from `https://openreview.net/forum?id=YicbFdNTTy`

Forbes, T., Linker, J., Chen, J., Cid, C., Kóta, J., Lee, M., ... others (2006). Cme theory and models: Report of working group d. *Space Science Reviews*, *123*(1), 251–302. Retrieved from `https://doi.org/10.1007/s11214-006-9019-8`

Garcia, H. A. (1994). Temperature and emission measure from goes soft x-ray measurements. *Solar Physics*, *154*(2), 275–308.

Grim, L. F. L., & Gradvohl, A. L. S. (2024). Solar flare forecasting based on magnetogram sequences learning with multiscale vision transformers and data augmentation techniques. *Solar Physics*, *299*(3), 33.

Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. (2022). Implementation paradigm for supervised flare forecasting studies: A deep learning application with video data. *Astronomy & Astrophysics*, *662*, A105.

Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. (2023). Operational solar flare forecasting via video-based deep learning. *Frontiers in Astronomy*

and Space Sciences, *9*, 1039805.

Hapgood, M. (2011). Towards a scientific understanding of the risk from extreme space weather. *Advances in Space Research*, *47*(12), 2059–2072.

Huang, X., Wang, H., Xu, L., Liu, J., Li, R., & Dai, X. (2018). Deep learning based solar flare forecasting model. i. results for line-of-sight magnetograms. *The Astrophysical Journal*, *856*(1), 7.

Kaneda, K., Wada, Y., Iida, T., Nishizuka, N., Kubo, Y., & Sugiura, K. (2022). Flare transformer: solar flare prediction using magnetograms and sunspot physical features. In *Proceedings of the asian conference on computer vision* (pp. 1488–1503).

Li, X., Zheng, Y., Wang, X., & Wang, L. (2020). Predicting solar flares using a novel deep convolutional neural network. *The Astrophysical Journal*, *891*(1), 10.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the ieee international conference on computer vision* (pp. 2980–2988).

Liu, H., Liu, C., Wang, J. T., & Wang, H. (2019). Predicting solar flares using a long short-term memory network. *The Astrophysical Journal*, *877*(2), 121.

Loshchilov, I., & Hutter, F. (2017). Sgdr: Stochastic gradient descent with warm restarts. In *International conference on learning representations (iclr)*. Retrieved from `https://openreview.net/forum?id=Skq89Scxx`

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations (iclr)*. Retrieved from `https://openreview.net/forum?id=Bkg6RiCqY7`

Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., & Ishii, M. (2018, may). Deep flare net (defn) model for solar flare prediction. *The Astrophysical Journal*, *858*(2), 113. Retrieved from `https://doi.org/10.3847/1538-4357/aab9a7` doi: 10.3847/1538-4357/aab9a7

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

NOAA Space Weather Prediction Center. (2024). *Solar flare classifications and radio blackout scales*. Retrieved from `https://www.swpc.noaa.gov/phenomena/solar-flares-radio-blackouts`

Park, E., Moon, Y.-J., Shin, S., Yi, K., Lim, D., Lee, H., & Shin, G. (2018). Application of the deep convolutional neural network to the forecast of solar flare occurrence using full-disk solar magnetograms. *The Astrophysical Journal*, *869*(2), 91.

Platts, J., Reale, M., Marsh, J., & Urban, C. (2022). Solar flare prediction with recurrent neural networks. *The Journal of the Astronautical Sciences*, *69*(5), 1421–1440.

Priest, E. R., & Forbes, T. (2002). The magnetic nature of solar flares. *The Astronomy and Astrophysics Review*, *10*(4), 313–377. Retrieved from `https://doi.org/10.1007/s001590100013`

Roy, S., Schmude, J., Lal, R., Gaur, V., Freitag, M., Kuehnert, J., ... Ramachandran, R. (2025). *Surya: Foundation model for heliophysics*. Retrieved from `https://arxiv.org/abs/2508.14112` (arXiv preprint)

Schou, J., Scherrer, P. H., Bush, R. I., Wachter, R., Couvidat, S., Rabello-Soares, M. C., ... others (2012). Design and ground calibration of the helioseismic and magnetic imager (hmi) instrument on the solar dynamics observatory (sdo). *Solar Physics*, *275*(1), 229–259.

Schrijver, C. J., Dobbins, R., Murtagh, W., & Petrinec, S. M. (2014). Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. *Space Weather*, *12*(7), 487–498.

Shibata, K., & Magara, T. (2011). Solar flares: magnetohydrodynamic processes. *Living Reviews in Solar Physics*, *8*(1), 1–99.

Solanki, S. K. (2003). Sunspots: an overview. *The Astronomy and Astrophysics Review*, *11*(2), 153–286. Retrieved from https://doi.org/10.1007/s00159-003-0018-4

Sun, P., Dai, W., Ding, W., Feng, S., Cui, Y., Liang, B., ... Yang, Y. (2022). Solar flare forecast using 3d convolutional neural networks. *The Astrophysical Journal*, *941*(1), 1.

Sun, Z., Bobra, M. G., Wang, X., Wang, Y., Sun, H., Gombosi, T., ... Hero, A. (2022). Predicting solar flares using cnn and lstm on two solar cycles of active region data. *The Astrophysical Journal*, *931*(2), 163.

Thomson, N. R., Rodger, C. J., & Clilverd, M. A. (2005). Large solar flares and their ionospheric d region enhancements. *Journal of Geophysical Research: Space Physics*, *110*(A6).

Tlatov, A., Abramov-Maximov, V., Borovik, V., & Opeikina, L. (2018). Evolution of solar active regions before large flares: Overview of the events of 2010–2017. *Geomagnetism and Aeronomy*, *58*(8), 1087–1096.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, X., Chen, Y., Toth, G., Manchester, W. B., Gombosi, T. I., Hero, A. O., ... Liu, Y. (2020). Predicting solar flares with machine learning: Investigating solar cycle dependence. *The Astrophysical Journal*, *895*(1), 3.

Yang, Y., Ni, Y. W., Chen, P., & Feng, X. S. (2025). Predicting solar flares using a convolutional neural network with extreme-ultraviolet images. *The Astrophysical Journal*, *985*(1), 104.