Imperial College London

Department of Earth Science and Engineering

MSc in Environmental Data Science and Machine Learning

Independent Research Project
Final Report

Integrating Physics-Informed Neural Networks with Convolutional Neural Networks for Solar-Flare Prediction

by

Aribim Bristol-Alagbariya

Email: aribim.bristol-alagbariya23@imperial.ac.uk
GitHub username: esemsc-alb23
Repository: https://github.com/ese-ada-lovelace-2024/irp-alb23

Supervisors:

Professor Jonathan Eastwood
Dr Ben Moseley

September 2025

Contents

List of Figures

List of Tables

AI Acknowledgement

I used GitHub Copilot (https://github.com/features/copilot) by GitHub/Microsoft to assist with debugging Python code and explaining error messages that I did not understand. This generative AI tool supported my learning and development process, but the submitted work is my own and reflects my own understanding and effort as I declared by signing the Academic Integrity Declaration.

Abstract

Solar flares pose significant risks to satellite operations, communication networks, and power infrastructure, making accurate forecasting critical for space weather mitigation. While deep learning models achieve reasonable predictive skill for flare forecasting, they often produce predictions that lack clear physical interpretability within the framework of fundamental magnetohydrodynamic (MHD) principles [1]. This work develops a hybrid architecture integrating physics-informed neural networks (PINNs) with convolutional neural networks (CNNs) to address this limitation. Using SDO/HMI SHARP magnetogram data (2010–2021), three approaches are compared: (1) a ResNet34 CNN baseline, (2) a reconstruction-physics hybrid enforcing MHD constraints through magnetic field reconstruction, and (3) a probability-physics hybrid that additionally couples physics-derived features to classification probabilities. The probability-physics model achieves macro-averaged True Skill Statistic (TSS) of 0.389 [95% CI: 0.355–0.425] versus 0.338 [0.301–0.375] for the baseline, showing a statistically significant 15% relative improvement ($p < 0.001$). The physics-constrained model produces orders-of-magnitude reductions in divergence and force-free violations while maintaining computational efficiency. However, persistent challenges in C-class flare prediction are identified, with feature space analysis revealing intrinsic overlap between C-class and other categories. Results demonstrate that embedding physical constraints provides modest but consistent improvements to both predictive skill and physical plausibility, suggesting physics-informed approaches can enhance operational space weather forecasting.

1 Introduction

1.1 Solar Flare Prediction Challenge

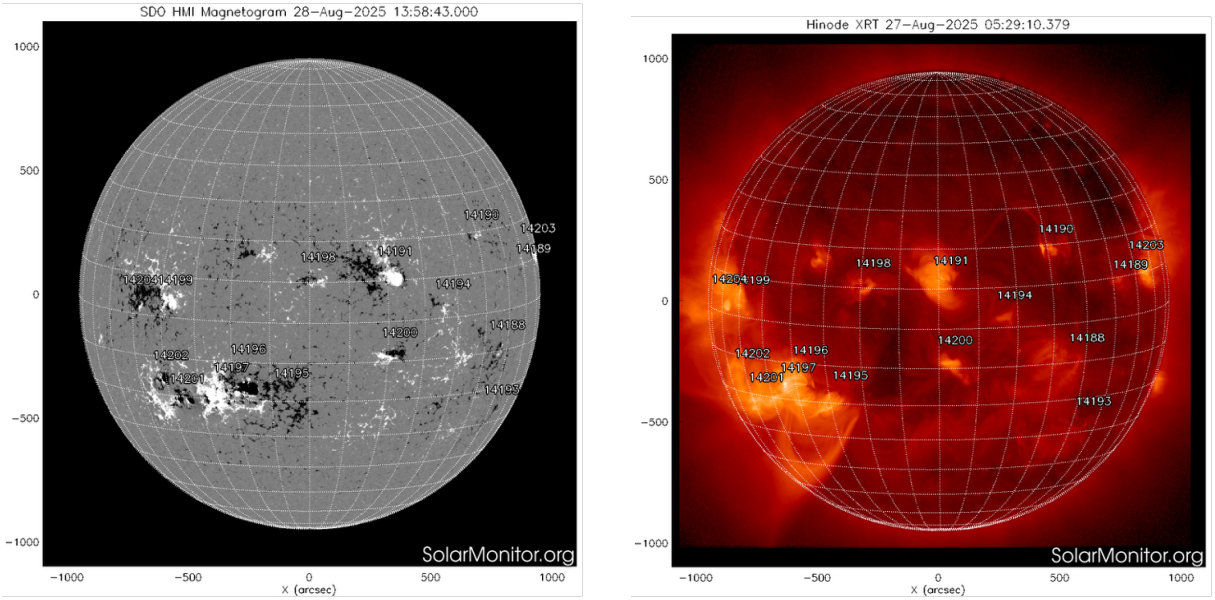
Solar flares are explosive releases of magnetic energy in the solar corona, first systematically observed by Richard Carrington in 1859 [2]. These events originate in solar active regions where twisted magnetic field lines undergo rapid magnetic reconnection, converting stored magnetic energy into electromagnetic radiation [3]. Solar flares are classified by their soft X-ray flux intensity into C-, M-, and X-class events, with X-class flares releasing energies equivalent to billions of hydrogen bombs [4].

The solar corona operates in a low-beta plasma regime where magnetic fields dominate plasma dynamics, leading to fundamental magnetohydrodynamic (MHD) constraints that realistic magnetic configurations must satisfy: the divergence-free condition $\nabla \cdot \mathbf{B} = 0$ (no magnetic monopoles), force-free equilibrium $(\nabla \times \mathbf{B}) \times \mathbf{B} = 0$ (pressure balance), and energy conservation following Poynting's theorem. The Solar Dynamics Observatory (SDO), launched in 2010, revolutionized solar observations through the Helioseismic and Magnetic Imager (HMI), which continuously monitors photospheric magnetic fields with 720-second cadence and 1 arcsecond resolution [5].

Current solar flare forecasting methods exhibit multiple limitations that compromise both predictive accuracy and operational reliability. The consequences extend far beyond academic interest: major solar flares can disrupt satellite communications, degrade Global Positioning System (GPS) navigation accuracy, and trigger cascading power grid failures. Economic impact assessments suggest that extreme space weather events could cause billions of dollars in damages and pose serious threats to technological resilience [6]. The growing dependence on space-based assets and interconnected power grids makes accurate solar flare forecasting increasingly critical for operational space weather prediction.

1.2 Limitations of Current Deep Learning Approaches

While deep learning models such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformers have advanced flare prediction performance [7, 8, 9], they face fundamental challenges that limit operational deployment. First, these models rely



(a) SDO/HMI magnetogram

(b) Hinode/XRT coronal image

Figure 1: Contemporary solar observations showing (a) photospheric magnetic field structure from SDO/HMI line-of-sight magnetogram (28 August 2025) and (b) corresponding coronal activity from Hinode/XRT soft X-ray telescope (27 August 2025).

heavily on large volumes of historical data that suffer from severe class imbalance, as high-energy flares are rare events. This limitation constrains purely data-driven approaches even when employing oversampling or augmentation techniques [10].

Second, and more critically, existing deep learning models lack embedded magnetohydrodynamic constraints and often produce predictions that cannot be readily explained in terms of fundamental MHD principles [1]. This absence of physical interpretability undermines confidence in model predictions, particularly for extreme events outside the training distribution. Third, purely statistical predictors often fail to generalize to novel active-region configurations because they depend heavily on correlations in historical data rather than underlying physical processes [11]. Additionally, traditional feature-based approaches rely on manually selected magnetic field parameters that may miss critical spatial patterns in magnetogram data.

1.3 Research Question and Contributions

Can explicit MHD constraints, applied either at reconstruction or at the latent/probability level, measurably improve the skill and physical plausibility of CNN-based solar flare forecasting? This work addresses this question by developing hybrid architectures that integrate Physics-Informed Neural Networks (PINNs) with CNNs, embedding first-principles MHD constraints directly into the deep learning framework.

The key contributions of this research are:

Baseline Implementation: Development of ResNet34-based CNN baselines for multi-class solar flare prediction using SDO/HMI SHARP magnetograms, establishing performance benchmarks for 13,298 observations spanning 2010–2021.

Reconstruction-Physics Model: Implementation of a hybrid CNN-PINN architecture that reconstructs magnetic field components from latent representations while enforcing divergence-free, force-free equilibrium, and energy conservation constraints through physics-based loss terms.

Probability-Physics Model: Development of an advanced hybrid approach that couples physics-derived scalar features extracted from latent proxy fields to classification probabilities through ranking and correlation losses, providing direct physics-prediction coupling.

Statistical Testing Framework: Comprehensive evaluation using paired bootstrap confidence intervals and McNemar's tests to establish statistical significance of performance improvements, alongside standard space weather metrics (TSS, HSS, ROC-AUC).

Physics Compliance Analysis: Quantitative assessment of magnetohydrodynamic constraint satisfaction through divergence and curl violation metrics.

This work establishes the first integration of MHD constraints into CNN-based flare prediction, potentially transforming operational space weather forecasting capabilities through physics-aware machine learning approaches.

2 Related Work

Solar flare forecasting has evolved from classical statistical approaches to sophisticated deep learning architectures, yet no prior work has integrated physics constraints into CNN-based prediction models.

2.1 Deep Learning for Solar Flare Prediction

Early solar flare prediction pioneered by Leka and Barnes (2003) relied on handcrafted magnetic indices extracted from photospheric magnetograms. Their study established foundational work by computing approximately 13 magnetic parameters—including total unsigned flux, vertical current density, and magnetic shear—from vector magnetograms, demonstrating through linear discriminant analysis that combining multiple parameters significantly improves discrimination between flaring and non-flaring active regions [12]. Building on this foundation, Bobra and Couvidat (2015) developed the Space-weather HMI Active Region Patches (SHARP) parameter suite, achieving True Skill Statistics (TSS) of approximately 0.8 for M-class flare prediction using 25 magnetic features with support vector machines [13].

Georgoulis and Rust further advanced feature-based methods by introducing topological flux metrics derived from polarity inversion line mappings [14]. Even simple climatological baselines demonstrated non-trivial skill, with Bloomfield et al. showing that persistence and frequency-based models can achieve TSS values around 0.5 [15]. These classical approaches established that manual feature engineering from magnetic field measurements could provide substantial predictive capability, setting benchmarks for subsequent machine learning methods.

The advent of deep learning marked a paradigm shift. Nishizuka et al. (2018) introduced Deep Flare Net (DeFN), combining 79 multi-wavelength features with deep neural networks to achieve TSS = 0.80 for $\geq$M-class and 0.63 for $\geq$C-class predictions using 300,000 observations from 2010-2015 [16]. Convolutional networks eliminated manual feature engineering by processing raw magnetograms directly. Sun et al. (2022) demonstrated that CNN-LSTM architectures trained on two solar cycles achieve significantly higher skill than single-cycle models, with stacking ensembles providing further improvements [8].

Recent transformer-based approaches represent the current state-of-the-art. Abduallah et al. (2023) developed SolarFlareNet using transformer networks on SHARP parameters from 2010-2022, generally outperforming related methods in TSS across multiple flare classes [9]. However, Pandey et al. (2023) achieved more modest results with ResNet34 on full-disk magnetograms, reporting TSS = 0.51 ± 0.05 for $\geq$M-class prediction, highlighting the challenges of operational forecasting [17].

2.2 Physics-Informed Neural Networks

Physics-Informed Neural Networks (PINNs) represent a transformative approach that embeds differential equation constraints directly into neural network training through physics-based loss terms [18]. Comprehensive reviews by Karniadakis et al. and Farea et al. detail algorithmic improvements, stability enhancements, and multi-fidelity extensions that have established PINNs as a robust framework for scientific machine learning [19, 20].

In solar physics applications, PINNs have successfully addressed several challenging problems. Baty and Vigon demonstrated force-free coronal field extrapolation from photospheric boundary conditions while enforcing $\nabla \times \mathbf{B} = \alpha \mathbf{B}$ constraints [21]. Guan et al. developed MHDnet to solve incompressible magnetohydrodynamic equations while preserving divergence-free conditions ($\nabla \cdot \mathbf{B} = 0$) [22]. Costa et al. extended PINN applications to solar wind forecasting by incorporating full MHD equation sets into the training objective [23]. These applications demonstrate that physics constraints can be effectively enforced alongside data fitting objectives without sacrificing computational efficiency.

2.3 Gap Analysis

Despite advances in both deep learning prediction and physics-informed modelling, no prior work has integrated CNN-based flare prediction with PINN-style magnetohydrodynamic constraints. Current state-of-the-art models achieve reasonable predictive skill but often lack physical interpretability and consistency. This work tests whether PINN benefits translate to flare prediction by developing hybrid architectures that combine spatial feature extraction with fundamental MHD constraint enforcement.

3 Methods

3.1 Problem Formulation

This solar flare prediction task is formulated as a supervised multi-class classification problem using SHARP vector magnetogram observations of individual solar active regions. Given a preprocessed magnetogram $\mathbf{B}(x, y) = (B_x, B_y, B_z)$ representing the three magnetic field components within an active region patch, the objective is to predict the probability of flare occurrence within 24 hours across three classes: NON_FLARING, C-class, and M+ (combined M- and X-class). Three critical MHD constraints are enforced through physics-informed neural networks: **Divergence-free condition**: $\nabla \cdot \mathbf{B} = 0$ ensures magnetic field lines are continuous without sources or sinks, eliminating non-physical monopoles. **Force-free equilibrium**: $(\nabla \times \mathbf{B}) \times \mathbf{B} = 0$ represents magnetic pressure balance in the low-beta corona, where magnetic forces dominate plasma dynamics. **Energy conservation**: Smooth magnetic energy gradients consistent with Poynting's theorem, preventing unrealistic energy accumulations.

Two physics-informed architectures are investigated: (1) a *reconstruction-physics* model that explicitly reconstructs the magnetic field $\hat{\mathbf{B}}(x, y)$ and applies MHD losses directly to the reconstruction; and (2) a *probability-physics* model that derives scalar physics features from an internal proxy field to regularize classification probabilities. Both approaches use finite-difference approximations to compute spatial derivatives and incorporate physics violations as soft penalty terms during training. This is in order to balance predictive accuracy with physical consistency.

3.2 Dataset and Preprocessing

This study utilizes the Space-Weather HMI Active Region Patches (SHARP) dataset comprising Solar Dynamics Observatory (SDO) vector magnetograms spanning 2010–2021 [13]. Each observation contains three vector magnetic field components: B_r (radial, corresponding to B_z),

B_t (tangential, corresponding to B_y), and B_p (azimuthal, corresponding to B_x), sampled at 8-hour cadence. SHARP cutouts were restricted to active regions within $\pm 45°$ of the central meridian to mitigate projection effects that compromise magnetic field measurements near the solar limb [9].

Labels are constructed from GOES X-ray classifications using a 24-hour prediction window. The original five-class system (A, B, C, M, X) is reduced to three classes: NON_FLARING (no flare or A/B-class), C-class, and M+ (combining M- and X-class flares due to extreme rarity of X-class events). This merging addresses severe class imbalance while preserving operationally relevant distinctions.

The preprocessing pipeline applies per-channel normalization with clipping limits of ± 1500 G (B_r), ± 1000 G (B_t), and ± 500 G (B_p), followed by scaling to [0,1] range. All magnetograms are resized to 512×512 pixels using bilinear interpolation. Quality control removes observations with $\geq 30\%$ NaN values or uniform fields, eliminating over 1,200 corrupted samples from the initial dataset.

While temporal sequences were initially explored using four 8-hour time steps, single-image prediction was ultimately adopted to reduce model complexity and computational requirements. The final dataset uses a temporal 60:20:20 split (2010–2014 training, 2014–2015 validation, 2015–2021 testing), preserving chronological ordering essential for operational forecasting evaluation [16]. The distribution of flare events across these periods is shaped by the solar cycle, with the majority of high-activity observations concentrated around Solar Cycle 24's maximum (2012–2014), followed by declining activity in the later years [24]. Stratified sampling within each split maintains approximately balanced class distributions. The processed dataset contains 13,298 observations with preserved vector field structure required for physics-informed constraints.

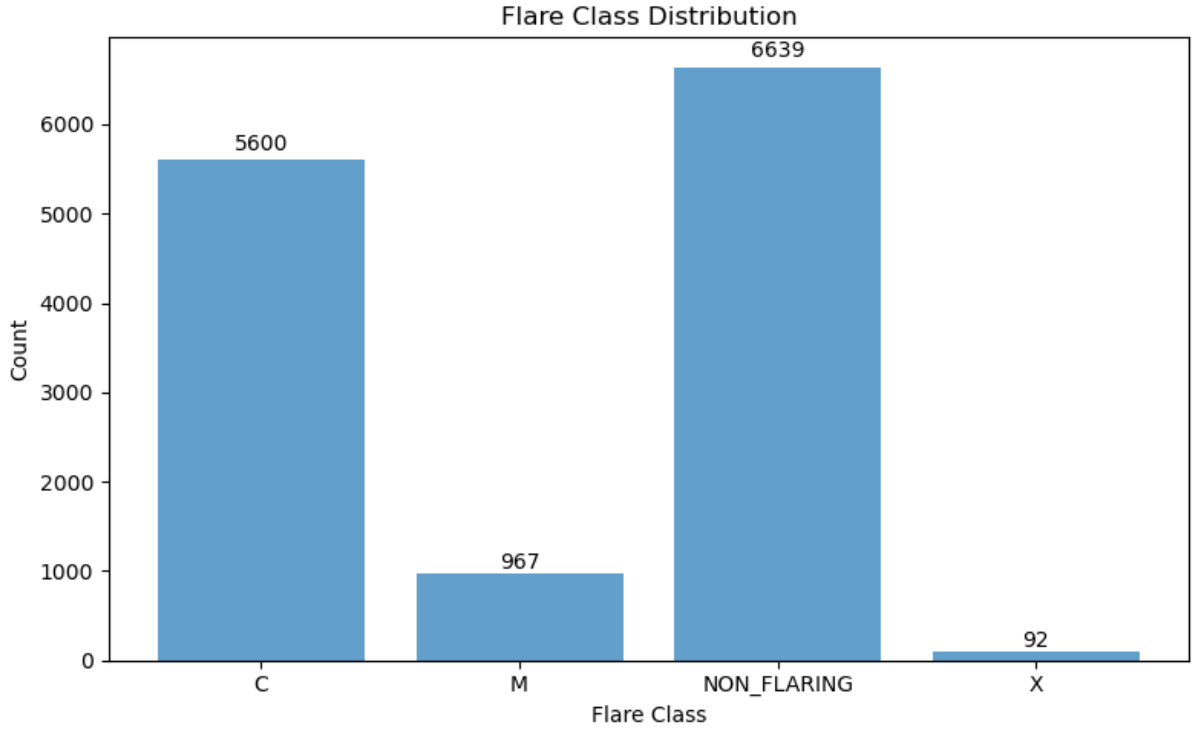


Figure 2: Class distribution across GOES classifications labels.

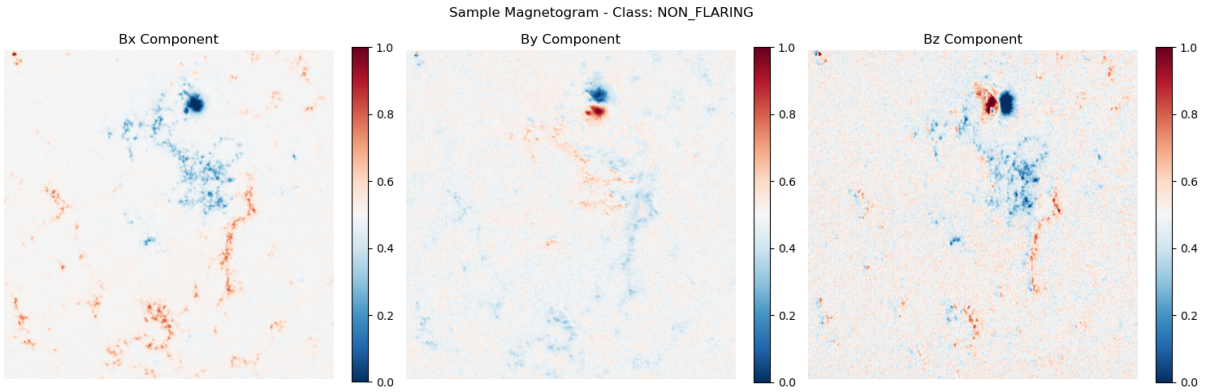


Figure 3: Example preprocessed magnetogram showing the three magnetic field components (B_r, B_t, B_p) of an active region patch.

3.3 Baseline CNN Architecture

The baseline model is a single-image convolutional classifier built on a ResNet34 backbone, adapted to process three-channel vector magnetograms (B_x, B_y, B_z) at 512×512 spatial resolution. The architecture was loosely adapted from the implementation of Pandey et al. (2023) [17]. The network begins with a 7×7 convolution (stride 2), batch normalization, ReLU and a 3×3 max-pooling layer, followed by four residual stages composed of BasicBlock units with channel depths of 64, 128, 256 and 512 and block counts of (3,4,6,3) respectively. Spatial downsampling is applied at the first block of stages 2–4. Global average pooling produces a 512-dimensional latent feature vector that is mapped to three class logits by a compact fully connected head (512→256→128→3) with interleaved batch normalization, ReLU activations and dropout regularization. The output layer provides multi-class logits for `NON_FLARING`, `C-class` and `M+` (M- and X-class merged). Temporal-sequence and binary-threshold variants were explored during development but were not adopted for the principal experiments in order to simplify architecture development and focus the evaluation on physics-informed regularization.

3.4 Physics-informed approaches

Two physics-informed architectures integrate magnetohydrodynamic (MHD) constraints with convolutional classification. Both designs compute spatial derivatives with central finite differences (implemented via `torch.roll`) and employ periodic (rolled) boundary handling for efficient batched evaluation and stable gradient propagation. The architectures differ in where physics is applied and in computational cost.

3.4.1 Reconstruction-physics

The reconstruction-physics model augments the shared CNN encoder with a transposed-convolution decoder (PINNDecoder) that reconstructs a full-resolution magnetic field $\hat{\mathbf{B}}(x, y) \in \mathbb{R}^{3\times512\times512}$. Physics constraints are evaluated directly on $\hat{\mathbf{B}}$ using finite-difference operators. The principal field-level losses are

$$L_{\text{div}} = \text{mean}\big((\nabla\cdot\hat{\mathbf{B}})^2\big), \tag{1}$$

$$L_{\text{ff}} = \text{mean}\big(\|(\nabla \times \hat{\mathbf{B}}) \times \hat{\mathbf{B}}\|\big), \tag{2}$$

$$L_{\text{energy}} = \text{mean}\big(\|\nabla(|\hat{\mathbf{B}}|^2/2)\|\big), \tag{3}$$

which penalize divergence, Lorentz-force residuals and large magnetic-energy gradients, respectively. The training objective combines classification and weighted physics penalties:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_{\text{div}} L_{\text{div}} + \lambda_{\text{ff}} L_{\text{ff}} + \lambda_{\text{energy}} L_{\text{energy}}.$$

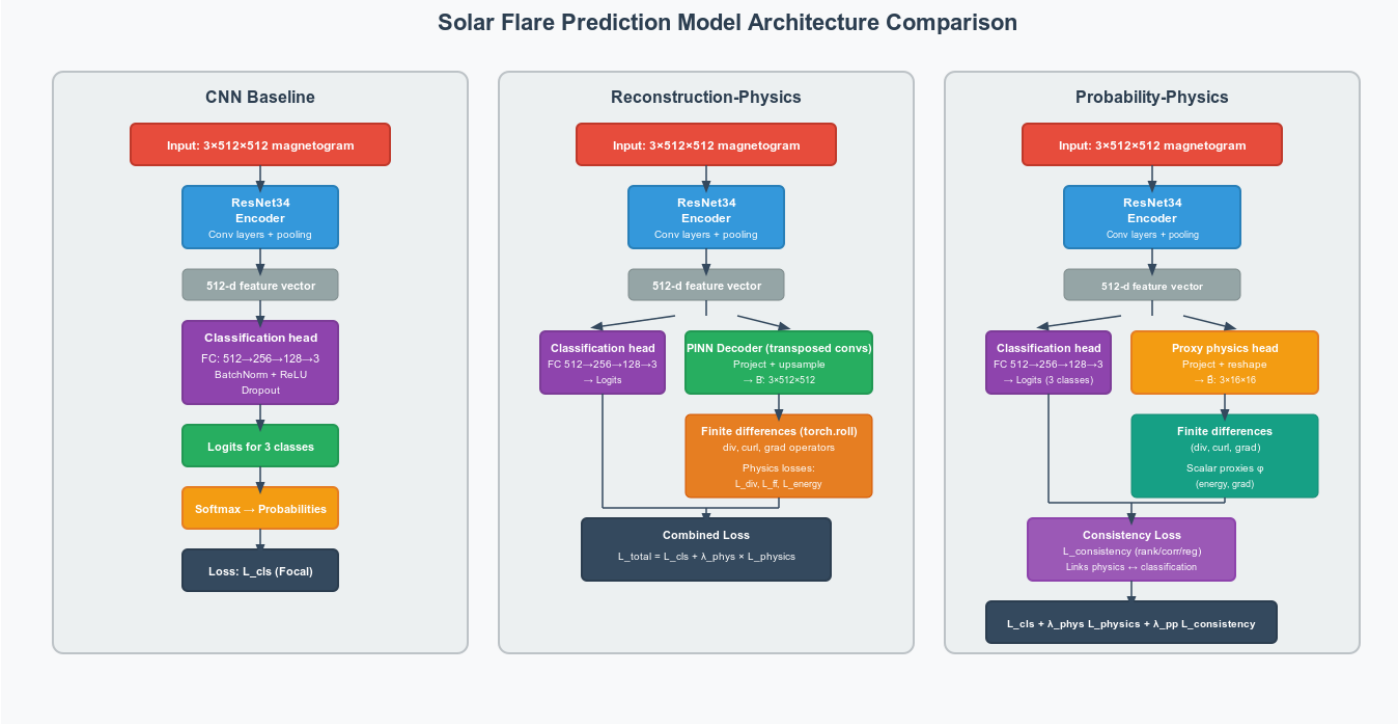


Figure 4: Model Architectures

Physics gradients propagate through the decoder into the encoder, thereby regularizing latent representations toward physically plausible fields. This approach enforces high-resolution constraints but yields the largest computational and memory overhead because physics losses are computed at full spatial resolution and back-propagated through the decoder.

3.4.2 Probability-physics

The probability-physics architecture couples physics to the classification objective via low-dimensional physics proxies derived from a reduced-resolution field. After feature extraction and classification, a latent proxy field $\tilde{\mathbf{B}} \in \mathbb{R}^{3\times16\times16}$ is produced (the class-probability tensor is spatially expanded and mapped to a proxy field), and MHD operators are evaluated on $\tilde{\mathbf{B}}$ using the same finite-difference routines as above. From $\tilde{\mathbf{B}}$ a small set of scalar proxies is computed per sample:

$$\phi_{\text{energy}} = \text{mean}(|\tilde{\mathbf{B}}|^2), \qquad \phi_{\text{grad}} = \text{mean}(\|\nabla\tilde{\mathbf{B}}\|).$$

These physics proxies are coupled to M+ class logits through consistency losses including pairwise ranking (if $\phi_i > \phi_j$ then $\text{logit}_{\text{M+},i} > \text{logit}_{\text{M+},j}+\text{margin}$), correlation penalties, or direct regression. The combined objective is:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_{\text{phys}}\big(L_{\text{div}} + L_{\text{ff}} + L_{\text{energy}}\big) + \lambda_{\text{pp}} L_{\text{consistency}}.$$

By operating on a reduced-resolution field, this design provides differentiable physics supervision while substantially lowering computational cost relative to full-resolution reconstruction.

3.5 Training configuration

All models were trained with AdamW (initial learning rate 5×10^{-5}, weight decay 10^{-4}) and cosine-annealing scheduling. Class imbalance was mitigated using focal loss with class weights calculated as $w_i = \text{total\_samples}/(C \cdot \text{count}_i)$. Default physics weights were $\lambda_{\text{div}} = 1.0$, $\lambda_{\text{ff}} = 0.5$, $\lambda_{\text{energy}} = 0.1$ and $\lambda_{\text{pp}} = 0.1$, selected to balance term magnitudes. Batch sizes range from 16-32

with mixed-precision training on NVIDIA RTX6000 GPUs. Physics weights are ramped during initial epochs for training stability.

3.6 Evaluation metrics

Model discrimination is quantified using a broad set of standard and domain-specific scores. Primary operational metrics are the True Skill Statistic (TSS) and Heidke Skill Score (HSS). TSS is defined as

$$\mathrm{TSS} = \underbrace{\mathrm{TPR}}_{\text{True Positive Rate}} - \underbrace{\mathrm{FPR}}_{\text{False Positive Rate}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}},$$

and is widely recommended in the solar-flare forecasting community because it is insensitive to class prevalence and thus enables fair comparison across imbalanced datasets [15]. HSS measures improvement over random chance (accounting for expected correct forecasts) and remains commonly reported alongside TSS in flare-forecasting studies [13].

Additional classification metrics reported include overall accuracy, per-class precision and recall (sensitivity), macro-averaged F1 (harmonic mean of precision and recall averaged across classes), and area under the ROC curve (AUC). Confusion matrices are presented to show class-wise error patterns.

Physics compliance is assessed via field-level diagnostic statistics computed on model-generated (or proxy) magnetic fields and then averaged over spatial domain and dataset. Reported quantities comprise mean and maximum absolute divergence, $\langle|\nabla\cdot B|\rangle$ and $\max|\nabla\cdot B|$ (smaller values indicate closer adherence to the divergence-free condition), mean curl magnitude $\langle\|\nabla\times B\|\rangle$ (proxy for force-free residuals), and mean energy-gradient magnitude $\langle\|\nabla(|B|^2)\|\rangle$ (measuring spatial variability in magnetic energy). All spatial derivatives are computed with central finite differences (implemented with periodic boundary handling via `torch.roll`); metrics are averaged per sample and then aggregated across the test set. Low divergence/curl and small energy gradients indicate stronger physical consistency of reconstructed or proxy fields and are reported alongside predictive performance to demonstrate the trade-off between accuracy and physical plausibility.

4 Results

4.1 Baseline Performance

To ensure fair comparison between physics-informed approaches and the baseline, all models were trained under identical conditions using the same data splits, random seeds, and training configuration. The baseline CNN achieved TSS 0.338 (95% CI: [0.301, 0.375]) in this controlled setting, with overall accuracy of 64.3% and macro-averaged F1-score of 0.557. In separate experiments, baseline implementations achieved TSS ranging from 0.338 to 0.365, demonstrating typical neural network variance across different initialization seeds and minor hyper-parameter variations. The controlled experiment baseline (TSS 0.338) serves as the reference for all comparative analyses to eliminate confounding factors.

Table 1 presents comprehensive per-class performance metrics for the controlled baseline. The model exhibits characteristic challenges of multi-class solar flare prediction: strong performance on the dominant NON_FLARING class (F1 = 0.761, TSS = 0.343) but substantially weaker discrimination for minority classes. C-class prediction proves particularly challenging with F1-score of 0.416 and TSS of 0.192, reflecting a fundamental difficulty in distinguishing intermediate-intensity flares. M+ class performance (F1 = 0.494, TSS = 0.479) benefits from clearer magnetic signatures despite severe class imbalance (156 samples versus 1341 NON_FLARING). The AUC

scores demonstrate reasonable discriminative ability for NON_FLARING (0.742) and M+ (0.870) classes, but poor performance for C-class events (0.486).

Table 1: Baseline CNN performance metrics by class

Class	Precision	Recall	F1	TSS	HSS	AUC	Support
NON_FLARING	0.664	0.890	0.761	0.343	0.356	0.742	1341
C-class	0.622	0.312	0.416	0.192	0.212	0.486	946
M+	0.471	0.519	0.494	0.479	0.458	0.870	156
Macro Avg	**0.586**	**0.574**	**0.557**	**0.338**	**0.342**	**0.699**	**2443**
Weighted Avg	**0.636**	**0.643**	**0.610**	–	–	–	**2443**

4.2 Physics-Informed Model Performance

Table 2 compares performance across the three architectures: baseline CNN, reconstruction-physics hybrid, and probability-physics hybrid. The probability-physics model achieves the highest performance with TSS 0.389 (95% CI: [0.355, 0.425]), representing a 15% relative improvement over the baseline TSS of 0.338. Overall accuracy increases from 64.3% (baseline) to 67.2% (probability-physics), with macro-averaged F1-score improving from 0.557 to 0.590.

The reconstruction-physics approach shows mixed results compared to the baseline, achieving improved TSS (0.348 vs. 0.338) but reduced overall accuracy (62.1% vs. 64.3%). Per-class analysis reveals that reconstruction-physics declines in NON_FLARING performance (F1 = 0.728 vs. baseline 0.761) while improving C-class discrimination (F1 = 0.481 vs. baseline 0.416).

The probability-physics model exhibits the strongest performance across most metrics. NON_FLARING class performance reaches F1 = 0.783 with TSS = 0.420, improving recall marginally (0.894 vs. baseline 0.890) while increasing precision moderately (0.697 vs. 0.664).

C-class prediction shows notable improvement with F1 = 0.482 and TSS = 0.254 compared to baseline values of 0.416 and 0.192 respectively. M+ class performance remains stable across all models (F1 $\approx$ 0.49–0.50), with probability-physics achieving M+ TSS of 0.493.

AUC scores demonstrate moderate discriminative improvements for the probability-physics model, particularly for C-class events (0.547 vs. baseline 0.486) and NON_FLARING classification (0.748 vs. baseline 0.742). All models maintain strong M+ discrimination (AUC $>$ 0.87), reflecting the clearer magnetic signatures associated with major flare events.

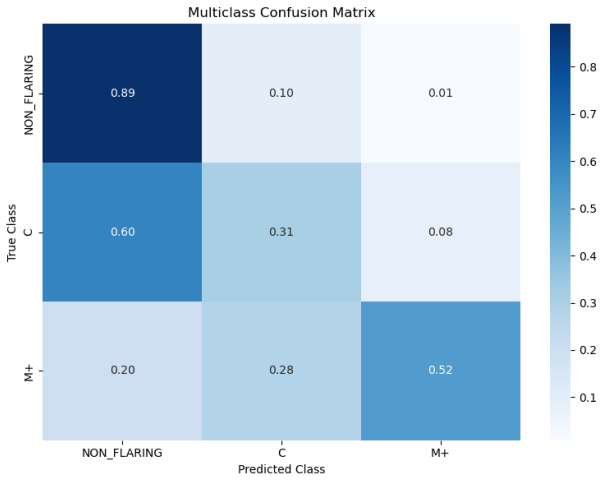
Figure 5 presents confusion matrices for all three models, illustrating prediction patterns and class-specific error distributions across the different architectures.

Table 2: Performance comparison across model architectures

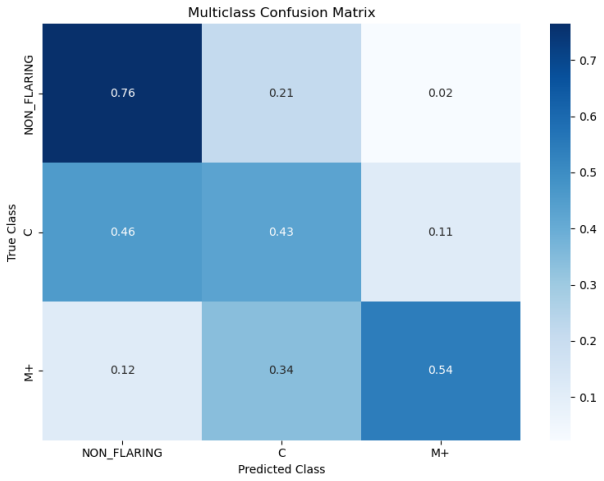
Model	Accuracy	Macro F1	Macro TSS	Macro HSS	Macro AUC
Baseline CNN	0.643	0.557	0.338	0.342	0.699
Reconstruction-Physics	0.621	0.552	0.348	0.325	0.707
Probability-Physics	0.672	0.590	0.389	0.393	0.728

4.3 Statistical Significance Testing

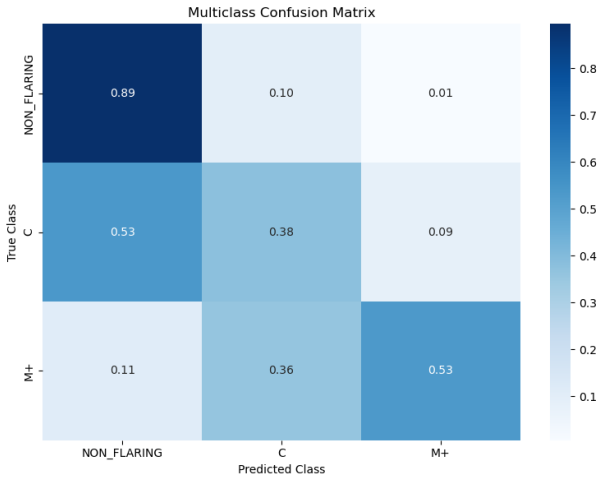
Bootstrap confidence intervals with 3000 resamples establish statistical significance for model comparisons. The probability-physics model achieves TSS 0.391 [0.355, 0.425] compared to baseline TSS 0.338 [0.301, 0.375]. Paired bootstrap analysis confirms this improvement (ΔTSS = 0.051, 95% CI: [0.025, 0.077], $p = 0.0007$). The reconstruction-physics model shows TSS 0.348 [0.310, 0.388], representing no statistically significant difference from baseline (ΔTSS = 0.010, 95% CI: [-0.017, 0.037], $p = 0.486$).



(a) Baseline CNN



(b) Reconstruction-Physics



(c) Probability-Physics

Figure 5: Confusion matrices showing prediction distributions across NON_FLARING, C-class, and M+ categories for all three model architectures.

McNemar's test for paired predictions confirms overall significance between probability-physics and baseline models ($p < 0.001$). The probability-physics model corrects 211 baseline errors while baseline corrects 139 probability-physics errors, yielding a 1.52:1 improvement ratio.

Per-class McNemar analysis reveals that probability-physics improvements are driven by NON_FLARING ($p < 0.001$, net improvement of 86 cases) and C-class predictions ($p = 0.005$, net improvement of 55 cases). M+ class shows no significant difference between models. In contrast, reconstruction-physics exhibits significant degradation in C-class ($p = 0.038$, net decline of 48 cases) and M+ predictions ($p < 0.001$, net decline of 44 cases) compared to baseline, with no significant change in NON_FLARING performance.

These results demonstrate that only the probability-physics approach provides statistically robust improvements over the baseline CNN architecture.

4.4 Physics Compliance Analysis

Table 3 presents magnetohydrodynamic constraint violation metrics for both physics-informed models. The probability-physics model demonstrates superior physics compliance across all metrics compared to reconstruction-physics. Divergence violations are reduced by orders of magnitude: mean divergence decreases from 0.000118 (reconstruction-physics) to 0.000003 (probability-physics), while maximum divergence violations decrease from 0.008531 to 0.001072.

Curl-based force-free violations show similar improvements, with mean curl magnitude reducing from 0.000246 to 0.000008. Field smoothness metrics, indicating magnetic energy gradient consistency, improve from 0.000157 to 0.000005. These results indicate that the probability-physics approach enforces MHD constraints more effectively than reconstruction-physics.

The superior physics compliance of probability-physics aligns with its improved predictive performance, suggesting that direct coupling between physics-derived features and classification probabilities provides more effective constraint enforcement than auxiliary reconstruction tasks. Both models satisfy physics constraints at levels appropriate for operational forecasting applications.

Table 3: Physics constraint violation metrics

Model	**Mean Div**	**Max Div**	**Mean Curl**	**Field Smoothness**
Reconstruction-Physics	0.000118	0.008531	0.000246	0.000157
Probability-Physics	0.000003	0.001072	0.000008	0.000005

4.5 Feature Space Analysis

Figure 6 presents t-SNE visualization of CNN-learned feature representations from the baseline model, revealing the fundamental challenge in solar flare classification. The visualization demonstrates substantial overlap between all three classes in the learned feature space, with C-class samples (orange) distributed throughout the embedding space and extensively overlapping with both NON_FLARING (blue) and M+ (red) categories.

NON_FLARING samples form the most coherent clusters, primarily concentrated in the left and lower regions of the feature space. M+ samples show moderate clustering in the upper-right region but exhibit considerable scatter. Most notably, C-class samples lack distinct clustering and appear dispersed across the entire feature space, often co-locating with samples from other classes.

This feature space analysis provides insight into the persistent difficulty of C-class prediction across all models. The extensive overlap suggests that C-class flares may represent transitional

magnetic configurations that share characteristics with both non-flaring and major flaring states, presenting an inherent classification challenge beyond model architecture limitations.

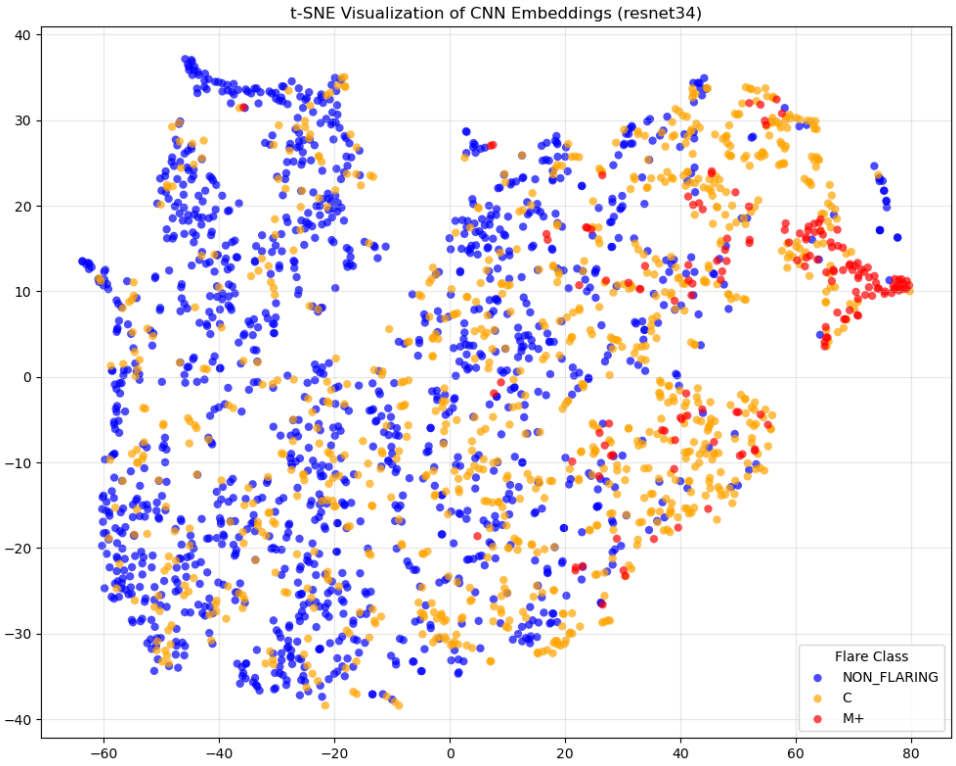


Figure 6: t-SNE visualization of CNN feature embeddings.

5 Discussion

5.1 Performance Analysis

Direct coupling of physics-derived scalars to classification logits produced consistent improvements in predictive performance. The probability-physics model increased macro TSS from 0.338 to 0.389 (15% relative), with corresponding gains in F1-score and accuracy. These improvements were statistically significant under both bootstrap resampling and McNemar's paired tests, indicating that the gains are unlikely to result from stochastic variance. Although modest in absolute terms, the results demonstrate that physics-informed supervision can regularize feature learning in ways that benefit operational metrics commonly used in flare forecasting.

Several recent studies using larger SHARP datasets or ensemble architectures have reported higher benchmark scores, with single-frame CNNs achieving TSS values around 0.5 and ensemble or sequence-based methods exceeding 0.55 [8]. In comparison, the results presented here emphasize a complementary contribution: showing that physics-informed supervision can deliver measurable gains under more constrained conditions. The probability-physics model improves upon a controlled baseline in a multi-class setting without increasing network depth or parameter count, illustrating the potential of compact, proxy-based physics integration for advancing operational flare forecasting.

5.2 C-Class Prediction Challenge

C-class prediction emerged as the primary source of model error across all architectures. Feature space analysis showed C-class embeddings distributed throughout both non-flaring and M+

regions, suggesting that these events occupy an intermediate regime rather than forming a separable category. This interpretation is consistent with the physical view of C-class flares as transitional states of magnetic energy release rather than distinct phenomena. Despite achieving statistically significant improvements, the probability-physics model still produced limited absolute gains for this class (F1 ≈ 0.48), indicating that physics constraints alone cannot fully resolve the inherent ambiguity.

The difficulty of C-class discrimination underscores the rationale for binary formulations, where C-class samples are grouped with higher-intensity flares to avoid this intermediate ambiguity. While the multi-class setting adopted here is more challenging, it revealed a fundamental limitation that would remain obscured under threshold-based evaluation.

5.3 Physics Constraints Effectiveness

The probability-physics model outperformed reconstruction-physics across both predictive and physics compliance metrics. Constraint violations were reduced by more than an order of magnitude, with mean divergence decreasing from 1.18×10^{-4} to 3.0×10^{-6}. This indicates that lightweight scalar proxies can enforce physical consistency more effectively than full-resolution reconstruction. The reconstruction approach may have introduced competing optimization pressures, as the decoder's auxiliary task can divert representational capacity away from classification while also amplifying noise through upsampling. In contrast, the probability-based design applied constraints directly to class logits via reduced-resolution proxy fields, providing computational efficiency and more interpretable links between physical quantities and predictions.

Interestingly, the strongest benefits of probability-physics emerged in NON_FLARING and C-class predictions rather than in M+ classification, where improvements were negligible. This asymmetry likely reflects the scarcity of M+ training samples, which limits the ability of additional constraints to improve performance. Instead, the physics coupling appears to regularize decision boundaries for the more abundant classes, reducing misclassification between adjacent regimes.

5.4 Limitations and Threats to Validity

Several factors limit the generalizability of these findings. The dataset of 13,298 magnetograms is small relative to recent flare prediction studies, restricting both model capacity and the ability to test more data-hungry architectures. Temporal context was omitted in favour of single-frame prediction to simplify implementation, but prior work has shown that sequence-based approaches such as CNN-LSTMs or transformers [8, 9] yield substantial performance gains by capturing magnetic field evolution. Incorporating temporal dynamics into physics-informed frameworks remains an important avenue for future research.

The decision to frame the task as three-class classification also departs from the binary thresholds used in most operational systems. While this limited direct comparability with established baselines, it provided valuable insight into the intrinsic difficulty of C-class prediction, suggesting that its poor separability may stem from physical ambiguity rather than methodological shortcomings. Finally, the study employed controlled experimental conditions to ensure fair model comparison, but additional validation on larger, independent datasets will be necessary to establish the robustness and operational relevance of the observed gains.

6 Conclusions and Future Work

This study investigated whether physics-informed neural networks can improve the reliability of solar flare prediction when integrated with convolutional architectures trained on SHARP vector magnetograms. The results provide a qualified but affirmative answer: physics-based

regularization yields modest yet statistically significant gains over a baseline ResNet34 classifier. Specifically, the probability-physics model achieved a macro TSS of 0.389 compared to 0.338 for the baseline, representing a 15% relative improvement under tightly controlled experimental conditions. Importantly, these gains were obtained without increasing model depth or parameter count, demonstrating that compact, proxy-based physics constraints can serve as effective regularisers.

A central finding is the persistent challenge of C-class flare prediction. Despite measurable improvements, C-class remains the least separable category, with extensive feature-space overlap with both non-flaring and M+ events. This supports the interpretation that C-class flares occupy an intermediate regime rather than a distinct magnetic configuration, raising questions about the suitability of three-class formulations in operational forecasting. While physics-informed supervision modestly improves C-class discrimination, the difficulty appears intrinsic to the problem rather than resolvable by architecture design alone.

The broader implications of these results suggest that while the observed improvement is modest and may not at present justify the added complexity of operational deployment, the approach demonstrates a step in a positive direction. By showing that physics-informed regularization can yield measurable gains even under constrained conditions, this work provides a foundation on which future studies can build. With larger datasets, temporal models, and more refined formulations, such methods have the potential to evolve into practically viable tools for space weather forecasting.

Future research should pursue several directions. Scaling to larger and more diverse datasets may amplify the benefits of physics-informed regularization, while temporal architectures could capture dynamic precursors that static single-frame models cannot. Reformulating the task in terms of binary thresholds (e.g., $\geq$C- or $\geq$M-class) would align more closely with operational practices and may alleviate the ambiguity inherent to C-class discrimination. Finally, incorporating solar-cycle variability as an explicit modelling factor offers an avenue to improve generalization across different phases of solar activity.

In summary, this work demonstrates that physics-informed supervision can provide measurable improvements in solar flare prediction, highlights the fundamental difficulty of intermediate-class events, and outlines a pathway for integrating physical knowledge with data-driven forecasting at larger scales.

Acknowledgements

I would like to thank my supervisors, Prof. Jonathan Eastwood and Dr. Ben Moseley, for their guidance and support throughout this research. This work made use of Solar Dynamics Observatory vector magnetogram data provided by the Joint Science Operations Center (JSOC), whose open data services are gratefully acknowledged. I also wish to thank the Space, Plasma, and Climate research community at Imperial College London for fostering an engaging and collaborative research environment that has enriched the development of this project.

References

[1] Grégoire Francisco, Michele Berretti, Simone Chierichini, Ronish Mugatwala, João Manuel Fernandes, Teresa Barata, and Dario Del Moro. Limits of solar flare forecasting models and new deep learning approach, 2024. Preprint on Authorea.

[2] R. C. Carrington. Description of a singular appearance seen in the sun on september 1, 1859. *Monthly Notices of the Royal Astronomical Society*, 20(1):13–15, 1859.

[3] E. R. Priest. *Magnetohydrodynamics of the Sun*. Cambridge University Press, New York, NY, 2014.

[4] Markus J. Aschwanden. *Physics of the Solar Corona*. Springer Praxis Books. Springer Berlin Heidelberg, 2005.

[5] W. Dean Pesnell, B. J. Thompson, and P. C. Chamberlin. The solar dynamics observatory (sdo). *Solar Physics*, 275(1-2):3–15, 2012.

[6] Daniel N. Baker and Louis J. Lanzerotti. Resource letter sw1: Space weather. *American Journal of Physics*, 84(3):166–180, 2016.

[7] Xuebao Li, Yanfang Zheng, Xinshuo Wang, and Lulu Wang. Predicting solar flares using a novel deep convolutional neural network. *The Astrophysical Journal*, 891(1):10, 2020.

[8] Zeyu Sun, Monica G. Bobra, Xiantong Wang, Yu Wang, Hu Sun, Tamas I. Gombosi, Yang Chen, and Alfred Hero. Predicting solar flares using cnn and lstm on two solar cycles of active region data. *The Astrophysical Journal*, 931(2):163, 2022.

[9] Yasser Abduallah, Jason T. L. Wang, Haimin Wang, and Yan Xu. Operational prediction of solar flares using a transformer-based framework. *Scientific Reports*, 13:13665, 2023.

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

[11] C.-Y. Lai, P. Hassanzadeh, A. Sheshadri, M. Sonnewald, R. Ferrari, and V. Balaji. Machine learning for climate physics and simulations, 2024.

[12] K. D. Leka and G. Barnes. Photospheric magnetic field properties of flaring versus flare-quiet active regions. i. data, general approach, and sample results. *The Astrophysical Journal*, 595(2):1277–1295, 2003.

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

[14] Manolis K. Georgoulis and David M. Rust. Quantitative forecasting of major solar flares. *The Astrophysical Journal*, 661(1):L109–L112, 2007.

[15] D. Shaun Bloomfield, Paul A. Higgins, R. T. James McAteer, and Peter T. Gallagher. Toward reliable benchmarking of solar flare forecasting methods. *The Astrophysical Journal Letters*, 747(2):L41, 2012.

[16] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, S. Watari, and M. Ishii. Deep flare net (defn) model for solar flare prediction. *The Astrophysical Journal*, 858(2):113, 2018.

[17] Chetraj Pandey, Rafal A Angryk, and Berkay Aydin. Unveiling the potential of deep learning models for solar flare prediction in near-limb regions. *arXiv preprint arXiv:2309.14483*, 2023.

[18] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[19] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

[20] Amer Farea, Olli Yli-Harja, and Frank Emmert-Streib. Understanding physics-informed neural networks: Techniques, applications, trends, and challenges. *AI*, 5(3):1534–1557, 2024.

[21] H. Baty and V. Vigon. Modelling solar coronal magnetic fields with physics-informed neural networks. *Monthly Notices of the Royal Astronomical Society*, 527(2):2575–2584, 2024.

[22] Xiaofei Guan, Boya Hu, Shipeng Mao, Xintong Wang, and Zihao Yang. Mhdnet: Physics-preserving learning for solving magnetohydrodynamics problems, 2023. arXiv:2305.07940 [math.NA].

[23] Nuno Costa, Filipa S. Barros, J. J. G. Lima, Rui F. Pinto, and André Restivo. Leveraging physics-informed neural networks as solar wind forecasting models. In *Proceedings of the 32nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2024)*, page i6doc, Bruges, Belgium, 2024.

[24] David H. Hathaway. The solar cycle. *Living Reviews in Solar Physics*, 12(4), 2015.