UNIVERSITÀ DI TRENTO

UNIVERSITÀ DELLA CALABRIA

UNIVERSITY OF TRENTO

Deep–Learning–Based Models for Solar Flare Forecasting

Author:
Elizabeth DORIA ROSALES

Supervisor:
Prof. Fabio LEPRETI

Co-supervisor:
Prof. Leonardo PRIMAVERA

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy

Doctorate of National Interest in Space Science and Technology

Physics Department

June, 2026

Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

"I press on toward the goal to win the prize for which God has called me heavenward in Christ Jesus."

Phi 3:14

"Keep your eyes on the prize, hold on

Ain't been to heaven, but I been told
Streets up there are paved with gold "

Eyes on the Prize

UNIVERSITY OF TRENTO

Abstract

Faculty Name
Physics Department

Doctor of Philosophy

Deep–Learning–Based Models for Solar Flare Forecasting

by Elizabeth DORIA ROSALES

One of the most important discoveries of the past century in solar physics was the detection of magnetic fields in sunspots by G. E. Hale, followed by the establishment of the existence of the much weaker global magnetic field of the Sun. Since then, the structure and polarity of magnetic fields in magnetically active regions (ARs) have been extensively investigated. Among the most prominent manifestations of solar magnetic activity are solar flares. Solar flares are localized explosive phenomena observed in the solar atmosphere. First observed in white light by R. Carrington, they are characterized by enhanced radiation across the electromagnetic spectrum. Flares are thought to originate from the rapid conversion of magnetic energy stored in stressed coronal fields and may be associated with other manifestations of solar activity, including solar energetic particle (SEP) events and, in some cases, coronal mass ejections (CMEs). Because flare electromagnetic radiation reaches the near–Earth environment within minutes, major flares can produce prompt ionospheric disturbances and affect radio communication, satellite navigation, and other technological systems that depend on the stability of the geospace environment. Despite their operational relevance, reliable flare forecasting remains challenging. Flare productivity depends on the nonlinear evolution of magnetic structures across multiple spatial and temporal scales, while the observational data are high–dimensional, heterogeneous, and strongly affected by the rarity of major events.

The Solar Dynamics Observatory (SDO) era has provided continuous, high–cadence observations of the solar atmosphere at multiple wavelengths, together with photospheric magnetic–field and Doppler measurements, and physical parameters derived from vector magnetic–field data. At the same time, recent developments in deep learning have made it possible to analyze such large observational datasets and extract relevant information. These developments motivate the central objective of this thesis: the investigation of deep–learning–based flare forecasting methods. Since solar flares are understood to be magnetically driven phenomena, active–region magnetograms constitute a natural starting point and a widely used source of predictive information. While many forecasting approaches rely on single–time representations of active–region magnetic field, this thesis focuses on the forecasting value of their temporal evolution and investigates how predictive performance is affected by progressively incorporating further sources of information, including continuum and extreme–ultraviolet (EUV) observations, as well as physical parameters derived from vector magnetic–field data.

The work is developed through four connected phases. First, convolutional autoencoders are used to learn compact representations of line–of–sight (LoS) magnetograms of ARs, with the aim of assessing whether the latent space preserves physically relevant information about AR magnetic morphology. Second, sequences of encoded LoS magnetograms are organized as temporal sequences and used as inputs to recurrent classifiers to predict whether an AR will produce an M– or X–class flare within the following hours, based on its prior evolution, thereby assessing the predictive power of both the learned representations and their temporal evolution.

Third, the forecasting framework is extended to a multimodal and multiwavelength setting that combines LoS magnetograms, continuum and EUV observations, and scalar physical parameters derived from vector magnetograms. In this case, the classification target is redefined as flares of class C5.0 or above, and the temporal analysis is restricted to 36 min of observations with a 2 h prediction window. Here, the focus shifts from the long–term evolution to short–term forecasting, where rapidly evolving photospheric and coronal conditions may be interpreted as preflare signatures. Finally, Layer–wise Relevance Propagation (LRP) is applied to interpret the predictions of this multimodal model and to examine how relevance is distributed across input modalities, as well as whether relevance allocation is associated with physically meaningful structures, such as polarity inversion lines (PILs) or regions of enhanced magnetic non–potentiality.

A first key innovation of this thesis is the use of compact neural representations learned through magnetogram reconstruction as inputs for flare forecasting. The results show that these representations can preserve both large–scale magnetic organization and smaller–scale magnetic structures, and that, when organized as temporal sequences, they also retain information on the temporal evolution of physical parameters derived from LoS magnetograms. Reconstruction objectives that explicitly constrain local magnetic–field gradients improve the recovery of gradient–sensitive physical diagnostics, indicating that the design of the learning objective affects the physical consistency of the latent representation. When LoS magnetogram representations are arranged as time series and used for forecasting, temporal modeling provides strong discrimination between flare–productive and non–flaring sequences, supporting the idea that flare–relevant information is encoded not only in the instantaneous magnetic configuration, but also in its evolution.

A second key innovation is the extension of the forecasting framework to a multimodal and multiwavelength setting. The multimodal forecasting experiments show that predictive performance improves as additional sources of observational and physical information are incorporated. Models based only on LoS magnetograms already capture useful predictive information, but the inclusion of continuum and EUV imagery provides complementary information about the photospheric and upper–solar–atmosphere evolution of ARs. The best performance is obtained when image–based representations are combined with physical parameters describing global magnetic properties.

The interpretability analysis indicates that the model decisions are often supported by localized relevance patterns in photospheric, transition–region, and coronal observations, with qualitative correspondence to physically meaningful photospheric magnetic structures, such as PILs, strong magnetic–field gradients, enhanced current helicity, and regions of increased free magnetic energy density, as well as to upper–solar–atmosphere structures traced by EUV emission, including localized brightenings and coronal loops. The relevance–allocation analysis further shows that image relevance generally accounts for most of the total relevance budget, while physical–parameter

relevance varies across classification outcomes and reflects different contributions from magnetic flux and current helicity. Misclassified cases reveal complementary failure modes: false positives are often linked to visually complex structures that resemble flare–productive morphology, while false negatives may correspond to regions in which physically relevant signatures are present but too localized or insufficiently emphasized by the learned image representation. These findings show that attribution methods are useful not only for interpreting successful predictions, but also for identifying the limitations and failure modes of deep–learning–based flare forecasting models.

This thesis contributes to solar flare forecasting by developing a unified deep–learning framework from data representation learning to interpretability. The results support the central conclusion that solar flare forecasting can benefit from frameworks able to integrate spatial structure, temporal evolution, multiwavelength observations, and physically motivated magnetic–field diagnostics. At the same time, the analysis highlights the importance of evaluating forecasting models not only in terms of predictive skill, but also in terms of calibration, robustness, physical consistency, and interpretability.

Acknowledgements

I would like to express my deepest gratitude to Prof. Vincenzo Carbone, my former supervisor, whose passing on January 21, 2025, left a profound void. I am sincerely thankful for his guidance and constant encouragement, and for the trust he placed in me and in my work. If I am here today, it is also because he never gave up in the face of a challenge and was always willing to learn new things and explore new scientific horizons.

I would like to thank Prof. Fabio Lepreti, who has been by my side from the very beginning, first together with Prof. Carbone as my co-supervisor and later as my supervisor. I am deeply grateful for his continuous support, especially during this last year and a half.

I would like to thank Prof. Leonardo Primavera, who kindly accepted to become my new co-supervisor and who has been doing a wonderful job ever since. I am grateful for his support and valuable contributions to my work.

I would like to thank Prof. Sergio Servidio, who has always shown genuine concern for the difficulties faced by PhD students. His support has been greatly appreciated.

My sincere thanks also go to Prof. Angelo Caramella and Dr. Emanuele Dinardo from the Università degli Studi di Napoli "Parthenope", and to Prof. Mariarosaria Falanga from the Università degli Studi di Salerno. We have worked together since the beginning, and part of the work presented in this thesis was carried out with their contribution. I am especially grateful for the many useful discussions and suggestions that enriched my research. Special thanks go to Emanuele, who has always been available and whose help has been essential to me in many ways.

I would also like to thank Prof. Consuelo Cid and Prof. Pablo Muñoz, who supervised me during my research period abroad at the Universidad de Alcalá in Madrid, as part of the Space Weather Group. My time there was extremely fruitful, both scientifically and personally. I had the opportunity to learn a great deal and to meet wonderful people, who made me feel truly welcome and at home. Thank you all from the bottom of my heart.

I would like to thank the administrative staff of the Università degli Studi di Trento for their great work in managing and supporting hundreds of students across Italy. Special thanks go to Gaia Fanelli for her infinite patience and kindness.

I would like to thank my dear friend Dr. Antonio Ielo. You have been a great help to me since my Bachelor's degree in Computer Science, and you still are. Thank you for your advice, your practical help, and your constant presence. But most of all, thank you for all the lunches, dinners, and walks on the bridge.

There are many more people I would like to thank than those listed so far. I have been fortunate to have always been surrounded by wonderful people who have enriched my life in so many ways. To every school teacher I have had, to every professor, to all my friends near and far, and to everyone who has helped me along the way: this is a dream come true, and it is also thanks to you.

Contents

Abstract iii

Acknowledgements vi

1 Introduction 1
- 1.1 Space Weather . . . 1
- 1.2 Impact on Earth and Human Technologies . . . 2
- 1.3 Thesis Objectives . . . 4
- 1.4 Thesis Structure . . . 5

2 Observational Aspects of Flare–Productive Active Regions 7
- 2.1 Sunspots Classification and Flare Productivity . . . 8
- 2.2 Formation and development of δ–spots . . . 11
- 2.3 Photospheric features . . . 13
  - 2.3.1 Strong–field, strong–gradient, highly–sheared PILs and magnetic channels . . . 13
  - 2.3.2 Photospheric Magnetic Complexity and Intermittency 14
  - 2.3.3 Current Neutralization and Multiscale Current Structure 16
- 2.4 Atmospheric and subsurface evolutions . . . 19
  - 2.4.1 Sigmoids and Filaments as Observational Signatures of Flux–Rope Topology . . . 19
  - 2.4.2 Non–thermal Broadening of EUV Spectral Lines Prior to Flares . . . 21
- 2.5 The Solar Dynamics Observatory (SDO) . . . 22
  - 2.5.1 The Helioseismic and Magnetic Imager (HMI) . . . 22
  - 2.5.2 The Atmospheric Imaging Assembly (AIA) . . . 25
  - 2.5.3 Space–weather HMI Active Region Patches (SHARPs) 26

3 Fundamentals of Neural Networks 29
- 3.1 From Perceptrons to Multilayer Networks . . . 30
- 3.2 Training Deep Neural Networks . . . 31
- 3.3 Convolutional Neural Networks (CNNs) . . . 33
- 3.4 Recurrent Neural Networks (RNNs) . . . 35
- 3.5 Attention Mechanisms . . . 36
- 3.6 Representation Learning . . . 38
- 3.7 Layer–wise Relevance Propagation (LRP) . . . 39
- 3.8 Deep Learning for Solar Flare Prediction . . . 41

4 Phase I: Representation Learning of Solar Active Regions 43
4.1 Introduction . . . 43
4.2 Data . . . 44
4.2.1 Data Sources . . . 44
4.2.2 Dataset Construction . . . 45
4.2.3 Train/Validation/Test Split . . . 47
4.2.4 Preprocessing . . . 47
4.3 Methodology . . . 49
4.3.1 Model . . . 49
4.3.2 Loss Functions . . . 52
4.3.3 Image Reconstruction Evaluation Tools . . . 56
4.3.4 Time–Series Evaluation Tools . . . 58
4.4 Results . . . 61
4.4.1 Model Selection . . . 61
4.4.2 Image Reconstruction Evaluation . . . 63
4.4.3 Temporal Evolution of Derived Physical Parameters . . . 71
4.4.4 Summary of Key Findings . . . 76
4.5 Conclusions . . . 77

5 Phase II: Spatio–Temporal Forecasting via Sequence Encoding 79
5.1 Introduction . . . 79
5.2 Data . . . 80
5.2.1 Data Sources and Dataset Construction . . . 80
5.2.2 Train/Validation/Test Split . . . 81
5.2.3 Sequence Extraction and Labeling . . . 81
5.2.4 Random Undersampling . . . 86
5.2.5 Preprocessing . . . 88
5.3 Methodology . . . 89
5.3.1 Model . . . 89
5.3.2 Evaluation Tools . . . 92
5.4 Results . . . 95
5.4.1 Model Selection . . . 96
5.4.2 Threshold Analysis . . . 101
5.4.3 Test–Set Forecasting Performance . . . 107
5.4.4 Discrimination and Probability Separation . . . 111
5.4.5 Calibration Analysis . . . 116
5.4.6 Summary of Findings . . . 121
5.5 Conclusion . . . 123

6 Phase III: Multimodal Spatio–Temporal Flare Forecasting 127
6.1 Introduction . . . 127
6.2 Data . . . 129
6.2.1 Data Sources . . . 129
6.2.2 Sequence Extraction and Labeling . . . 130
6.2.3 Train/Validation/Test Split . . . 133
6.2.4 Preprocessing . . . 134
6.3 Methodology . . . 135
6.3.1 Model . . . 135

6.3.2 Evaluation Tools . . . 137
6.4 Results . . . 139
Classifier (Base Model: Only LoS Magnetogram) . . . 140
Classifier (Base Model + Continuum and EUV) . . . 141
Classifier (Base Model + Continuum and EUV + Physical Quantities) . . . 141
6.4.1 Summary of Key Findings . . . 144
6.5 Conclusions . . . 145

7 Physical Interpretation of Multimodal Solar Flare Forecasting 148
7.1 Introduction . . . 148
7.2 Data and Model . . . 149
7.2.1 Dataset and Preprocessing . . . 149
7.2.2 Model . . . 150
7.3 Methodology for Interpretability Analysis . . . 152
7.3.1 Attribution Method . . . 152
7.3.2 Physics–based Diagnostic Maps . . . 155
7.3.3 Definition of Quantitative Attribution Metrics . . . 156
7.3.4 Selection of Samples for Interpretation . . . 159
7.4 Results . . . 161
7.4.1 Quantitative Attribution Analysis . . . 161
7.4.2 Physical–Parameter Distributions by Classification Outcome . . . 167
7.4.3 Case Studies . . . 168
True Positive: HARP 5298 . . . 169
False Positive: HARP 3730 . . . 173
True Negative: HARP 2779 . . . 173
False Negative: HARP 2809 . . . 176
7.4.4 Summary of Key Findings . . . 176
7.5 Conclusions . . . 179

8 Conclusions 182
8.1 Main Findings and Contributions . . . 182
8.2 Future Perspectives . . . 186
8.3 Final Remarks . . . 188

A Data Availability 189

B Temporal Distribution of the Datasets 190
B.1 Active–Region–Centered Dataset: Chapters 4 and 5 . . . 190
B.2 Event–Centered Dataset: Chapter 6 . . . 190

Bibliography 192

List of Figures

2.1 Temporal and latitudinal evolution of sunspot activity from 1874 to 2016. The upper panel shows the butterfly diagram of sunspot area in equal–area latitude strips, expressed as a percentage of each strip area, illustrating the systematic equatorward migration of sunspot emergence during successive solar cycles, in accordance with Spörer's law. The lower panel shows the corresponding disk–integrated average daily sunspot area, expressed as a percentage of the visible solar hemisphere, with solar cycle numbers indicated near each activity maximum. 9

4.1 Schematic representation of the CAE architecture used for LoS magnetogram reconstruction. The encoder progressively reduces the spatial resolution through strided convolutions while increasing the number of feature channels. The decoder mirrors this process using transposed convolutions to reconstruct the input LoS magnetogram. 51

4.2 Comparison of the radially averaged Fourier power spectra for the original LoS magnetograms and the reconstructions obtained with the selected models. The two examples correspond to HARP 377 (upper panels), observed on 2011.02.15 at 02:00 during an X2.2 flare, and HARP 8906 (lower panels), observed on 2022.12.21 at 12:48 and not associated with a flare. The left panels show the full spectra, while the right panels show a zoom of the high–k range. The dashed horizontal line marks the power value at $k = 32.5$, which is used as a reference for the high–k range shown in the zoomed panels. 65

4.3 Representative reconstruction example for HARP 377, observed on 2011.02.15 at 02:00 during an X2.2 flare. The reconstruction is obtained with the finite–difference gradient–regularized model. The panels show the original and reconstructed LoS magnetograms, polarity–separated absolute residual maps, and the pixel–value distributions. 67

4.4 Representative reconstruction example for HARP 8906, observed on 2022.12.21 at 12:48 and not associated with a flare. The reconstruction is obtained with the finite–difference gradient–regularized model. The panels show the original and reconstructed LoS magnetograms, polarity–separated absolute residual maps, and the pixel–value distributions. 68

4.5 Local reconstruction comparison for HARP 377, observed on 2011.02.15 at 02:00 during an X2.2 flare. The red ellipse in the full LoS magnetogram identifies the selected region. The figure compares the selected original region with the reconstructions obtained using the reconstruction–only baseline model, the Sobel–regularized model, and the finite–difference gradient–regularized model, together with their absolute deviation maps. 69
4.6 Local reconstruction comparison for HARP 8906, observed on 2022.12.21 at 12:48 and not associated with a flare. The red ellipse in the full LoS magnetogram identifies the selected region. The figure compares the selected original region with the reconstructions obtained using the reconstruction–only baseline model, the Sobel–regularized model, and the finite–difference gradient–regularized model, together with their absolute deviation maps. 70
4.7 Temporal evolution of the derived physical parameters for HARP 377. The panels show the total unsigned LoS magnetic flux Φ, the gradient–weighted neutral line length R, and the mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$. Black curves indicate the parameters computed from the original magnetograms, while colored curves indicate the parameters computed from magnetograms reconstructed with the selected models. Vertical red lines indicate flares of class $\geq$ C5.0. Gray shaded regions indicate missing observations, while blue shaded regions indicate intervals with no PIL formation. 74
4.8 Temporal evolution of the derived physical parameters for HARP 8589. The panels show the total unsigned LoS magnetic flux Φ, the gradient–weighted neutral line length R, and the mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$. Black curves indicate the parameters computed from the original magnetograms, while colored curves indicate the parameters computed from magnetograms reconstructed with the selected models. Vertical red lines indicate flares of class $\geq$ C5.0. Gray shaded regions indicate missing observations, while blue shaded regions indicate intervals with no PIL formation. 75

5.1 Example schematic representation of the sequence selection and labeling procedure. Each candidate sequence is defined by an observational window of duration T, sampled with cadence Δt, followed by a forecasting window of duration T used for label assignment. A sequence is labeled as positive if at least one M– or X–class flare occurs during the forecasting window, and as negative if no flare occurs during either the observational or forecasting window. Sequences corresponding to a decay in flare activity or to cases with only weak flares during the forecasting window are excluded from the classification task. 84

5.2 Example schematic representation of the treatment of missing records during sequence construction. When a record is missing, a placeholder is created. A candidate sequence is discarded when a placeholder record falls within the observational window, because the model input must be complete and regularly sampled. If a placeholder falls within the forecasting window of a candidate sequence with a complete observational window, the sequence can still be retained, since the forecasting window is used only for label assignment. . . . 85
5.3 Schematic representation of the forecasting model architecture. Each magnetogram in the input sequence is processed independently by a shared frozen convolutional encoder with attention pooling, producing a 512–dimensional feature vector for each time step. The resulting sequence of feature vectors is passed to an LSTM layer, and the final hidden state is used by a fully connected classifier to produce the binary prediction. . 91
5.4 Validation–set model selection for the static–undersampling strategy. Each column corresponds to one temporal configuration, and each cell represents a combination of LSTM hidden size and learning rate. The upper row reports the minimum validation loss reached by each configuration, while the lower row reports the corresponding validation *TSS* evaluated at the same epoch using the fixed threshold $\tau = 0.5$. 97
5.5 Validation–set model selection for the dynamic–undersampling strategy. Each column corresponds to one temporal configuration, and each cell represents a combination of LSTM hidden size and learning rate. The upper row reports the minimum validation loss reached by each configuration, while the lower row reports the corresponding validation *TSS* evaluated at the same epoch using the fixed threshold $\tau = 0.5$. 98
5.6 Validation *TSS* as a function of the decision threshold for the selected static–undersampling models. The solid curves show the validation *TSS* obtained across threshold values, the colored vertical dashed lines indicate the *TSS*–optimal thresholds selected on the validation set, and the gray vertical dashed line marks the default threshold ($\tau = 0.5$). 102
5.7 Validation *TSS* as a function of the decision threshold for the selected dynamic–undersampling models. The solid curves show the validation *TSS* obtained across threshold values, the colored vertical dashed lines indicate the *TSS*–optimal thresholds selected on the validation set, and the gray vertical dashed line marks the default threshold ($\tau = 0.5$). 102
5.8 Test–set ROC curves for the selected static–undersampling models. The diagonal dashed line indicates the no–skill reference, and the markers indicate the operating points obtained using validation–optimized thresholds. 112

5.9 Test–set ROC curves for the selected dynamic–undersampling models. The diagonal dashed line indicates the no–skill reference, and the markers indicate the operating points obtained using validation–optimized thresholds. 112
5.10 Test–set predicted–probability distributions for the selected static– and dynamic–undersampling models. Columns correspond to the four temporal configurations, while rows correspond to the undersampling strategy. Each panel shows the probability distributions for negative and positive sequences, with the vertical dashed line indicating the validation–optimized threshold. 114
5.11 Test–set reliability diagram for the selected static–undersampling models. The dashed diagonal line indicates perfect calibration. Each point represents a non–empty probability bin, with the x–axis corresponding to the mean predicted probability and the y–axis corresponding to the observed positive fraction. The histograms show the distribution of predicted probabilities for each temporal configuration. 118
5.12 Test–set reliability diagram for the selected dynamic–undersampling models. The dashed diagonal line indicates perfect calibration. Each point represents a non–empty probability bin, with the x–axis corresponding to the mean predicted probability and the y–axis corresponding to the observed positive fraction. The histograms show the distribution of predicted probabilities for each temporal configuration. 119

6.1 Examples of active region extraction and quality control. Each panel shows, from left to right, SDO/HMI LoS magnetograms, SDO/HMI continuum intensity images, and SDO/AIA EUV images in the 193 and 304 Å channels, all reprojected to the AIA 304 Å WCS and taken 2 h prior to the flare peak time. Red boxes delineate the full extent of the active region extracted from the SHARP data product, while blue boxes indicate the reference dimensions of the adopted cutouts. (a) NOAA 11158 (HARP 377): HARP original bounding–box size was 377×744 pixels, exceeding 512 pixels in one dimension but not the other. Consequently, the final cutout extent along each axis was defined as the greater of 512 pixels and the original dimension, resulting in a cutout of 512×744 pixels. (b) NOAA 12192 (HARP 4698): the bounding box enclosing the active region crossed the solar limb within the look–back window; consequently, the corresponding sequence was excluded from the dataset. 132

6.2 Overview of the proposed solar flare prediction framework. Left: a convolutional autoencoder is trained on individual multi–wavelength solar images (x) to learn latent representations (z), enabling unsupervised feature extraction. Right: a supervised prediction pipeline integrates learned image representation and physical features. Latent representations (z_0, z_1, z_2) are obtained by applying the encoder independently to each image in the sequence, while hand–engineered physical parameters (f_0, f_1, f_2) undergo a linear transformation to match the dimensionality of the learned features, allowing the two modalities to be summed and fused into a context–aware representation. The fused feature sequence is then processed by a recurrent sequence encoder, followed by a classifier that outputs the predicted probability of a flare event. 136
6.3 Classifier performance across different input modalities. Columns represent models trained on: (1) LoS magnetograms only; (2) LoS magnetograms with additional continuum and AIA image channels; and (3) full multimodal input including SHARP physical parameters. Top row: predicted probability distributions for positive and negative samples. Second row: calibration curves comparing predicted probabilities with observed outcome frequencies. Third row: ROC curves and *ROC–AUC* values, showing model discrimination capability. Bottom row: *Precision*, *Recall*, and *F1–score* as functions of the decision threshold. Performance improves consistently as additional data modalities are incorporated. 143

7.1 Schematic illustration of the forward prediction pass and backward relevance propagation used in the interpretability analysis. Red arrows indicate the forward flow from the image sequence and physical–parameter time series to the prediction, while blue arrows indicate the backward LRP flow from the prediction back to the input–level relevance tensors. The dashed box marks the shared latent representation where learned image and physical–parameter features are combined. 154
7.2 Distribution of fractional relevance–allocation metrics grouped by classification outcome for the MCAF model. The left panel compares the fraction of total absolute relevance assigned to the image and physical–context branches. The right panel compares the internal fraction of physical–context relevance assigned to Φ_{tot} and $h_{c,\text{tot}}$. 161
7.3 Distribution of the standardized total unsigned magnetic flux, $z_{\Phi_{\text{tot}}}$, and total unsigned current helicity, $z_{h_{c,\text{tot}}}$, grouped by classification outcome. The dashed horizontal line indicates the test–set mean after standardization. 168

7.4 Final input frame for the four representative ARs used in the case–study analysis. Columns correspond to the *TP*, *FP*, *TN*, and *FN* cases. Rows show HMI B_{LoS}, HMI I_c, AIA 304 Å, and AIA 193 Å. In the B_{LoS} row, the yellow contours indicate the polarity inversion lines (PILs). In the I_c row, red and blue contours indicate positive and negative magnetic polarity, respectively. 170
7.5 Physics–based diagnostic maps for the four representative ARs used in the case–study analysis. Columns correspond to the *TP*, *FP*, *TN*, and *FN* cases. Rows show the LoS magnetic field, the free magnetic energy density ρ_{free}, the vertical current helicity h_c, and the horizontal gradient magnitude $|\nabla B_{\mathrm{LoS}}|$. . . . 171
7.6 LRP relevance maps for the *TP* case, HARP 5298, associated with an M3.2 flare. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å. 172
7.7 LRP relevance maps for the *FP* case, HARP 3730. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å. 174
7.8 LRP relevance maps for the *TN* case, HARP 2779. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å. 175
7.9 LRP relevance maps for the *FN* case, HARP 2809, associated with an M1.0 flare. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å. 177

B.1 Temporal distribution of the ARs used in Chapters 4 and 5 across the training, validation, and test partitions. The upper panel shows the monthly sunspot number, while the lower panel shows the observation epoch of each AR. The dashed vertical lines indicate the approximate boundaries of Solar Cycles 24 and 25. 191
B.2 Temporal distribution of the Chapter 6 positive and negative sequences across the training, validation, and test partitions. Marker shape distinguishes positive and negative samples, while color indicates the data partition. The upper panel shows the monthly sunspot number for temporal reference. 191

List of Tables

4.1 Summary of the final dataset, including the number of retained ARs, the total number of SHARP records, and the number of weak ($\leq$ C–class) and strong ($\geq$ M–class) flare events. Percentages in the flare columns are computed relative to the total number of flares. 46

4.2 Summary of the active–region–level train, validation, and test split, including the number of ARs, the total number of records, and the number of weak ($\leq$ C–class) and strong ($\geq$ M–class) flare events in each subset. 48

4.3 Percentage distribution of the active–region–level train, validation, and test split. Percentages in the total–records column are computed relative to the full number of records, while percentages in the flare columns are computed relative to the total number of flare events within each subset. 48

4.4 Validation *MSE* for the candidate autoencoder configurations. Rows indicate the primary reconstruction loss and the auxiliary loss, while columns report the auxiliary–loss weight λ. The baseline column reports configurations trained without an auxiliary loss. 62

4.5 Selected autoencoder configurations retained for the final test–set evaluation. The table reports the best reconstruction–only baseline, the best finite–difference gradient–regularized model, and the best Sobel–regularized model according to validation MSE. 63

4.6 Image–level reconstruction metrics computed on the test set for the selected autoencoder configurations. The table reports pixel–wise reconstruction errors, structural similarity, and distributional agreement between the original and reconstructed LoS magnetograms. 64

4.7 Test–set temporal–evolution metrics for the derived physical parameters computed from the original and reconstructed LoS magnetograms. The table reports results for the selected models: the reconstruction–only baseline, the Sobel–regularized model, and the finite–difference gradient–regularized model. Lower values are better for *MAE*, *sMAE*, *DTW distance*, and *sDTW distance*, while higher values are better for ρ. 73

5.1 Definition of sequence class labels for the binary classification task. Positive samples are denoted by "+", and negative samples by "−". Blank entries correspond to a decay in flare activity, while "?" entries correspond to sequences leading to weak–flare activity; both cases are excluded under the labeling scheme adopted for the present task. 82
5.2 Number and percentage of positive and negative sequences generated for each temporal configuration. A temporal configuration is defined by the duration of the observational window in hours, the duration of the forecasting window in hours, and the sampling cadence of the input sequence in minutes. . 86
5.3 Best models for each temporal configuration and undersampling strategy. The selected model is the one achieving the minimum validation loss. Other classification metrics on the validation set are reported using the fixed threshold $\tau = 0.5$. . 100
5.4 Threshold sensitivity for the selected models. The threshold interval indicates the range over which the validation *TSS* remains within 0.01 of its maximum value. 104
5.5 Validation–set performance of the selected models at the *TSS*–optimal threshold. The optimal threshold is selected by maximizing validation *TSS*; when multiple thresholds achieve the same maximum *TSS*, the lowest value among the tied optima is selected. All remaining metrics are computed at the selected threshold. 105
5.6 Effect of threshold optimization on validation–set performance for the selected models. The default threshold is $\tau = 0.5$, while the optimized threshold is selected by maximizing validation *TSS*. Differences are computed as optimized–threshold performance minus default–threshold performance. 106
5.7 Confusion–matrix counts and classification metrics computed on the test set for the selected models using thresholds optimized on the validation set. 108
5.8 Direct comparison between static and dynamic undersampling on the test set. For each temporal configuration, metric differences are computed as dynamic minus static for *TSS*, *Recall*, and *Specificity*. 109
5.9 Ranking of the selected models according to test–set *TSS*. *Recall* and *Specificity* are reported to show the *Recall–Specificity* trade–off obtained at each validation–optimized threshold. . . 110
5.10 Threshold–independent discrimination and probability–separation metrics computed on the test set for the selected models. *ROC–AUC* and *PR–AUC* quantify discrimination, while the mean and median probability differences summarize the separation between truly positive and truly negative sequences. The sample columns report the number of positive and negative sequences for each temporal configuration for reference. 115

5.11 Test–set calibration metrics for the selected models. The *Brier score* measures the mean squared error of the predicted probabilities, while *ECE* and *MCE* quantify the expected and maximum calibration gaps across probability bins, respectively. Lower values indicate better calibration. 121

6.1 Summary of the dataset composition grouped by GOES flare class. The flare category includes events above the C5.0 threshold, with each valid flare event contributing a single observational sequence. The no–flare category denotes samples with no associated flaring activity within the prediction window. . 133
6.2 Summary of the convolutional autoencoder architecture used to compress individual four–channel input images into compact latent representations. The encoder consists of five convolutional blocks with increasing channel depth, while the decoder mirrors this structure using transposed convolutions to reconstruct the input image. Activation functions include ReLU for intermediate layers and *tanh* at the output of the decoder. . 138
6.3 Architecture of the sequence classifier used to predict flare occurrence from latent representations. The model consists of a context mixing layer that projects the physical parameters into the same dimensional space as the latent image representations, followed by a two–layer LSTM for sequence processing and a fully connected classifier. A final sigmoid activation outputs the probability of a flare event within the prediction window. 138
6.4 Confusion matrix values for all models evaluated in this study. All quantities are computed on the held–out test set. 142
6.5 Performance metrics for all models evaluated in this study. All quantities are computed on the held–out test set. 144

7.1 Confusion–matrix counts for the model configurations used in the interpretability analysis. 152
7.2 Layer–specific propagation rules adopted for the modified forecasting model used in this chapter. The rules are assigned according to the role of each layer in order to preserve numerical stability and produce interpretable relevance maps. 153
7.3 Consensus misclassifications in the test set. The table includes only samples misclassified by all three model configurations. The quantities p_M, p_{MCA}, and p_{MCAF} denote the predicted flare probabilities produced by model M (LoS magnetogram–only), model MCA (LoS magnetogram + continuum + EUV), and model MCAF (LoS magnetogram + continuum + EUV + physical parameters), respectively. 160
7.4 Representative samples selected for the case–study interpretation. The selected examples include one *TP*, one *TN*, one *FP*, and one *FN* case. The reported values are the predicted flare probabilities from the three model configurations. 160

7.5 Summary statistics of the relevance–allocation metrics grouped by classification outcome for the MCAF model. Values are reported as median $[Q_1, Q_3]$, except for N, which gives the number of samples in each group. The metric $\Delta_{\mathrm{img-ctx}}^{|\cdot|}$ is defined as the base–10 logarithmic difference between the image and physical-context relevance fractions. 163
7.6 Summary statistics of the total absolute relevance assigned to each image channel by the MCAF model over the full test set. 165
7.7 Summary statistics of the total absolute relevance assigned by the MCAF model to the image sequence at each time step over the full test set. 165
7.8 Summary statistics of the total absolute relevance assigned by the MCAF model to the physical–parameter time series at each time step over the full test set. 165
7.9 Summary statistics of the positive and negative relevance–fraction metrics grouped by classification outcome for the MCAF model. Values are reported as median $[Q_1, Q_3]$, except for N, which gives the number of samples in each group. The metrics f^+ and f^- quantify the fractions of total relevance magnitude associated with positive and negative relevance contributions, respectively. 166

List of Abbreviations

AARP	AIA Active Region Patch
Adam	adaptive moment estimation
AIA	Atmospheric Imaging Assembly
AR	active region
BBSO	Big Bear Solar Observatory
BPTT	backpropagation through time
CAE	convolutional autoencoder
CCD	charge–coupled device
CCF	cross–correlation function
CEA	Cylindrical Equal–Area
CME	coronal mass ejection
CNN	convolutional neural network
DEM	differential emission measure
DTW	dynamic time warping
ECE	expected calibration error
EVE	Extreme Ultraviolet Variability Experiment
EUV	extreme–ultraviolet
FD	finite difference
FN	false negative
FNR	false negative rate
FP	false positive
FPR	false positive rate
FUV	far–ultraviolet
GAN	generative adversarial network
GD	gradient descent
GNSS	Global Navigation Satellite System
GOES	Geostationary Operational Environmental Satellite
Grad–CAM	Gradient–weighted Class Activation Mapping
HARP	HMI Active Region Patch
HF	high frequency
HMI	Helioseismic and Magnetic Imager
HS	hidden size
IG	Integrated Gradients
IQR	interquartile range
JSOC	Joint Science Operations Center
KL	Kullback–Leibler
LDM	latent diffusion model
LoS	line–of–sight
LRP	Layer–wise Relevance Propagation

LSTM	long short–term memory
MAE	mean absolute error
MCE	maximum calibration error
MDI	Michelson Doppler Imager
MHD	magnetohydrodynamic
MSE	mean squared error
MSSIM	mean structural similarity index
NOAA	National Oceanic and Atmospheric Administration
NRT	near–real–time
PIL	polarity inversion line
PR	precision–recall
PR–AUC	area under the precision–recall curve
ReLU	rectified linear unit
RNN	recurrent neural network
ROC	receiver operating characteristic
ROC–AUC	area under the receiver operating characteristic curve
SDO	Solar Dynamics Observatory
SEE	single–event effect
SEP	solar energetic particle
SGD	stochastic gradient descent
SHAP	Shapley Additive exPlanations
SHARP	Space–weather HMI Active Region Patch
SMOTE	synthetic minority over–sampling technique
SOHO	Solar and Heliospheric Observatory
SOT	Solar Optical Telescope
SSIM	structural similarity index
sMAE	standardized mean absolute error
TN	true negative
TNR	true negative rate
TP	true positive
TPR	true positive rate
TSS	true skill statistic
UV	ultraviolet
VAE	variational autoencoder
VFISV	Very Fast Inversion of the Stokes Vector
VQGAN	vector–quantized generative adversarial network
VQ–VAE	vector–quantized variational autoencoder
WCS	world coordinate system

List of Symbols

cm	centimeter	length
m	meter	length
km	kilometer	length
s	second	time
min	minute	time
h	hour	time
ns	nanosecond	time
Hz	hertz	frequency
MHz	megahertz	frequency
nm	nanometer	wavelength
Å	angstrom	wavelength
W	watt	power
J	joule	energy
erg	erg	energy
$\mathrm{W\,m^{-2}}$	watt per square meter	irradiance / flux
G	gauss	magnetic field strength
$\mathrm{km\,s^{-1}}$	kilometer per second	velocity
pixel	pixel	image size / spatial sampling
arcsec	arcsecond	angular resolution
degree	degree	angle

To my family
To my friend
To my love

Chapter 1

Introduction

"There are more things in heaven and earth, Horatio,
Than are dreamt of in your philosophy. "

William Shakespeare
Hamlet

1.1 Space Weather

Space weather refers to variations in the space and near–Earth environment that can affect technological systems, human activities, and society. The dominant driver of space weather is the Sun, and more specifically the evolution, storage, and release of magnetic energy in the solar corona. Magnetic fields generated by plasma motions in the solar interior emerge through the photosphere into the corona, where they may form complex magnetic configurations capable of storing substantial free magnetic energy. Magnetic reconnection is thought to be the process through which this stored energy can be rapidly converted into plasma heating, particle acceleration, and bulk motion, giving rise to solar flares that may, in some cases, be accompanied by coronal mass ejections (CMEs) that propagate into interplanetary space (Priest and Forbes, 2002). Together, these processes define the fundamental physical chain linking solar activity to geospace response. The principal Earth and near–Earth systems affected are the electromagnetic fields within the solid body of Earth, the particle–radiation environments in Earth's atmosphere and magnetosphere, and the density, composition, and dynamics of the high–altitude atmosphere, particularly the coupled ionosphere–thermosphere system.

Space weather has become increasingly relevant in the past few decades, as modern society has become more reliant on advanced technological systems whose operation depends on the stability of the near–Earth environment. Historically, this vulnerability can be traced to telegraph disruption in the nineteenth century (Barlow, 1849), early telephone vulnerabilities (Preece, 1894), and power–grid disturbances in the twentieth century (McNish, 1940), to the far broader technological exposure of the satellite era. Forecasting

adverse space–weather conditions is not only a matter of scientific understanding, but also part of the broader effort to anticipate and mitigate risk in technologically advanced societies whose energy, communication, navigation, timing, and space–based infrastructures are sensitive to variations in the near–Earth environment.

Within this general framework, solar flares occupy an important role because they are among the fastest and most direct solar drivers of geospace disturbances (Schwenn, 2006). Flares result from magnetic reconnection in the solar corona, where electrons are accelerated to high energies and some propagate downward toward denser plasma near the solar surface. Their interaction with this dense plasma produces the intense X–ray and extreme–ultraviolet (EUV) emission that characterizes a flare (Schrijver, 2009). From the perspective of solar–terrestrial coupling, this is particularly significant because flare radiation reaches Earth essentially at the speed of light and can therefore modify the dayside ionosphere on timescales of minutes. This prompt radiative pathway distinguishes flares from slower, though often highly geoeffective, drivers such as CMEs, whose propagation from the Sun to Earth generally requires many hours to days.

1.2 Impact on Earth and Human Technologies

Space weather affects Earth and human technologies through multiple physical coupling mechanisms, but these mechanisms are not equally relevant to all solar drivers. In the broadest sense, space weather can disrupt electric power systems, radio communications, satellite navigation, digital electronics and spacecraft operations. Some of these effects are primarily associated with geomagnetic storms and CME–driven magnetospheric forcing, whereas others arise more directly from flare radiation. It is therefore useful to distinguish between the full spectrum of space–weather impacts and those that are most immediately and physically linked to flares. This distinction does not diminish the importance of the broader space–weather hazard context; rather, it identifies the specific impact mechanisms that make solar flare prediction operationally relevant within space–weather forecasting and mitigation.

The most direct flare–related coupling mechanism with technological consequences is the disturbance of the ionosphere by flare–enhanced X–ray and EUV radiation. During large X–class solar flares, solar irradiance can increase by three to four orders of magnitude in the X–ray band (0.1–10 nm) and by more than a factor of two in the EUV (10–120 nm) and far–ultraviolet (FUV; 120–200 nm) bands (Tsurutani et al., 2005; Woods et al., 2005; Scherliess, 2016). This enhanced irradiance produces additional photoionization on the dayside of Earth and increases the ionospheric plasma density within minutes. The altitude of the response depends on wavelength: hard X–rays (< 1 nm) penetrate to the lower ionosphere, mainly the D region (60–90 km), while soft X–rays (1–10 nm) and EUV radiation reach higher layers, including the E region (90–150 km) and the F region, above approximately 150 km (Rishbeth and Garriott, 1969; Zolesi and Cander, 2013; Buzás et al., 2023). In

large events, X–ray ionization can increase D–region electron density by one order of magnitude or more, while EUV–driven ionization can enhance electron density by approximately 50–200% in the E region and 10–100% in the F region (Liu et al., 2024).

These ionospheric changes modify radio–wave propagation almost immediately, particularly for high–frequency (HF) radio communication, where long–distance communication relies on signals being refracted by the ionosphere back toward Earth and, in some cases, propagating over multiple hops between the ionosphere and Earth's surface (Davies, 1990). Enhanced D–region ionization increases the absorption of HF radio waves (3–30 MHz) and can produce partial or total short–wave fadeouts lasting from tens of minutes to hours, thereby raising the effective lowest usable frequency for HF communication (George and Bradley, 1973; Rishbeth and Garriott, 1969; Davies, 1990; Tsurutani et al., 2009; Barta et al., 2019; Barta et al., 2022). This is operationally important because HF radio remains relevant for aviation air–to–ground communication, military and governmental communication systems, and maritime communication services. Flare–enhanced photoionization at different ionospheric altitudes can also alter global ionospheric electrodynamics and current systems, and affect magnetosphere–ionosphere coupling. Flare effects are therefore not confined to the ionosphere where the radiation is absorbed, but can extend through geospace via electrodynamic coupling (Liu et al., 2021a; Liu et al., 2021b).

Another important impact channel concerns Global Navigation Satellite Systems (GNSS). GNSS services are best known for their use in navigation, but they also provide precise timing services that support financial transactions, telecommunications, transport networks, and other critical infrastructures (Hapgood, 2018; Cannon et al., 2013). GNSS signals must pass through the ionosphere before reaching ground–based receivers, and the presence of free electrons modifies the signal propagation time relative to propagation in vacuum. Even small signal delays of only tens of nanoseconds can lead to position errors of several meters. In addition to propagation delays, ionospheric irregularities can produce amplitude and phase scintillation, degrading signal quality and, in severe cases, causing receivers to lose phase lock (Hapgood, 2018; Cannon et al., 2013). Flare–enhanced irradiance in the X–ray and EUV bands provides one of the fastest mechanisms for modifying dayside ionospheric plasma density, rapidly altering the propagation medium through which satellite navigation and timing signals travel (Yasyukevich et al., 2018). Flares may also be accompanied by solar radio bursts at GNSS frequencies, which can increase the receiver noise background, reduce the carrier–to–noise ratio, and degrade tracking and positioning performance (Sreeja, Aquino, and Jong, 2013; Sato et al., 2019; Paula et al., 2022).

Solar flares can be associated with solar energetic particle (SEP) events; when the particle flux reaching near–Earth space becomes sufficiently enhanced, solar radiation storms can occur, substantially increasing the radiation environment in near–Earth space and, in some cases, allowing energetic particles to penetrate into Earth's atmosphere (Reames, 2013; Reames, 2021; Ryan, Lockwood, and Debrunner, 2000). One important technological

consequence of this enhanced radiation environment is its effect on digital electronics present in a wide range of space–based technologies (Normand, 1996). High–energy particles can interact with semiconductor devices and deposit charge in sensitive regions of a circuit, producing single–event effects (SEEs). The best–known example is the single–event upset, in which the deposited charge changes the value of a stored bit, for example from 0 to 1 or vice versa. Other forms of SEE can be more severe, including latch–up, where a device becomes locked in a particular state, and events that cause permanent component damage. The vulnerability of modern electronics to these effects has increased as digital devices have become more miniaturized. As the feature sizes of semiconductor components decrease, smaller amounts of deposited charge can be sufficient to alter or disrupt device operation. Consequently, enhanced particle radiation associated with solar eruptive activity represents an important space–weather hazard for spacecraft electronics, avionics, and other technologies exposed to elevated radiation environments (Normand, 1996; Dyer et al., 2000).

Other consequences of space weather also remain important. For example, accelerated orbital decay in low–Earth–orbit satellites is primarily caused by solar–forcing–induced variations in thermospheric density (Walterscheid, 1989; Lei et al., 2013; Nwankwo, Chakrabarti, and Weigel, 2015). Enhanced solar EUV and X–ray irradiance, including that associated with flares, can heat and expand the upper atmosphere on timescales of hours, while geomagnetic storms can produce additional thermospheric heating through magnetospheric and ionospheric energy injection (Buonsanto, 1999; Chen et al., 2012; Kutiev et al., 2013). The resulting increase in atmospheric density enhances drag on satellites and debris in low Earth orbit, accelerating orbital decay rates and affecting orbit prediction, conjunction analysis, and reentry forecasting (Klinkrad, 1996; Doornbos and Klinkrad, 2006; Xu et al., 2011).

Solar flares therefore occupy a distinctive position within the broader space–weather hazard context. Their most direct consequences arise from the rapid modification of the ionosphere, with downstream effects on radio propagation and GNSS–based navigation and timing services, while associated energetic particles and upper–atmospheric responses broaden their relevance to spacecraft operations and the near–Earth orbital environment. This combination of fast onset, technological relevance, and limited warning time provides the operational motivation for improving major–flare forecasting.

1.3 Thesis Objectives

The advent of space–based solar observations has transformed solar physics by providing continuous, high–resolution, multiwavelength coverage of the Sun. At the same time, advances in deep learning and computational hardware now make it possible to process these large observational datasets and extract information that is difficult to capture with manually defined features alone. In this context, the general objective of this thesis is to investigate how deep–learning methods can be used to improve solar flare forecasting

from observations of solar active regions (ARs) and to support the physical interpretation of model predictions.

The first objective of this thesis is to examine whether compact neural representations of active–region line–of–sight (LoS) magnetograms can preserve information relevant to the magnetic structure of flare–productive ARs. This objective is addressed by training convolutional autoencoders to reconstruct LoS magnetograms and by evaluating whether different learning objectives affect the quality and physical consistency of the reconstructed ARs. In this context, reconstruction is used not as an end in itself, but as a way to assess whether the latent representation retains meaningful information about the photospheric magnetic field.

The second objective is to determine whether learned LoS magnetogram representations can be extended to the temporal domain for flare forecasting. Solar flares are not determined only by the instantaneous configuration of an AR, but also by its evolution. For this reason, this thesis investigates recurrent neural-network models for temporal flare forecasting, using sequences of learned magnetogram representations to assess whether active–region evolution improves the prediction of major flares.

The third objective is to move from single–modality forecasting toward a more complete representation of the solar atmosphere. Since flare productivity depends on physical processes that involve multiple atmospheric layers and magnetic–field properties, this thesis develops a multimodal forecasting framework that combines photospheric magnetic–field observation, continuum intensity observations, coronal EUV emission, and physical parameters derived from vector magnetograms. This objective addresses whether complementary observational channels and expert–defined physical quantities can improve the predictive skill of deep–learning flare–forecasting models.

The fourth objective is to examine the physical interpretability of the forecasting models. High predictive performance alone is not sufficient in a scientific context if the model behaves as an opaque classifier. Therefore, this thesis applies attribution methods to investigate which input modalities, spatial regions, and physical parameters contribute most strongly to the model predictions. The goal is to assess whether the model decisions are consistent with known physical properties of flare–productive ARs.

These objectives define a progression from representation learning, to temporal forecasting, to multimodal prediction, and finally to physical interpretation. This thesis therefore aims to contribute not only to the development of more accurate flare–forecasting models, but also to the broader question of how deep–learning methods can be made more physically informative in solar physics.

1.4 Thesis Structure

This thesis is organized as follows:

Chapter 1 introduces the space–weather context of this thesis, the relevance of solar flares for Earth and human technologies, and the main objectives of this work.

Chapter 2 presents the observational solar–physics background required for this thesis. It reviews the properties of flare–productive ARs, the magnetic conditions associated with flare productivity, and the observational signatures used to characterize active–region evolution.

Chapter 3 introduces the deep–learning concepts used throughout this thesis. It describes the neural–network architectures and methodological tools relevant to the following chapters, including convolutional neural networks, recurrent models, attention mechanisms, representation learning, and interpretability methods.

Chapter 4 presents the first experimental phase of this thesis, focused on representation learning of solar ARs. A convolutional autoencoder is trained to reconstruct LoS magnetograms, and different training objectives are evaluated in terms of both image reconstruction quality and the preservation of physically meaningful magnetic–field properties.

Chapter 5 extends the representation–learning framework to flare forecasting. Encoded LoS magnetogram sequences are used as inputs to a long short–term memory (LSTM)–based forecasting model in order to assess whether the temporal evolution of active–region magnetic structure improves the prediction of major flares.

Chapter 6 develops a multimodal spatio–temporal forecasting model. The model combines LoS magnetograms, continuum intensity observations, coronal EUV images, and physical parameters derived from vector magnetic–field data. The chapter evaluates how the progressive inclusion of complementary information affects short–term forecasting performance.

Chapter 7 focuses on the physical interpretation of the multimodal forecasting model. Layer–wise Relevance Propagation (LRP) is used to analyze the contribution of different modalities, image regions, and physical parameters to the model predictions. The results are discussed in relation to known physical diagnostics of flare–productive ARs.

Finally, the concluding chapter summarizes the main findings of this thesis, discusses their implications for solar flare forecasting and interpretable deep learning in solar physics, and outlines possible directions for future work.

Chapter 2

Observational Aspects of Flare–Productive Active Regions

"I have no data yet. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."

Arthur Conan Doyle
The Adventures of Sherlock Holmes
(A Scandal in Bohemia)

Solar flares are among the most energetic transient events in the solar system, releasing up to about 10^{32} erg of magnetic energy over timescales of seconds to minutes. According to the classical stress–and–release model, flares occur when magnetic energy accumulated in the coronal field due to photospheric motions is suddenly released via magnetic reconnection (Priest and Forbes, 2002). This process rapidly converts stored magnetic energy into kinetic and thermal forms (Benz, 2017). Although it is widely accepted that the coronal magnetic field serves as the primary reservoir of flare energy (Forbes, 2000; Schrijver, 2009), the precise mechanisms governing energy storage and release remain poorly understood.

Radiative signatures of flares extend across the entire electromagnetic spectrum, from radio wavelengths to γ rays. Operationally, they are classified according to their peak soft X–ray flux in the Geostationary Operational Environmental Satellite (GOES) 1–8 Å band into the logarithmic classes A, B, C, M, and X, corresponding to flux levels of 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, and $10^{-4}\,\mathrm{W\,m^{-2}}$ at Earth, respectively. Each class is further subdivided from 1.0 to 9.9, except for X–class flares, which can exceed 10.0; for example, X1.2 and M3.4 correspond to $1.2 \times 10^{-4}\,\mathrm{W\,m^{-2}}$ and $3.4 \times 10^{-5}\,\mathrm{W\,m^{-2}}$, respectively.

Flares also exhibit a characteristic temporal evolution, commonly describe in terms of a precursor phase, an impulsive phase, and a more extended gradual phase, each associated with distinct radiative signatures and physical processes. During the impulsive phase of solar flares, energetic particles are accelerated; some of these particles may escape from the Sun along magnetic field lines originating close to the flare site, producing impulsive

solar energetic particle (SEP) events in interplanetary space (Reames, 1999; Reames, 2013). Large–scale expulsions of magnetized coronal plasma into interplanetary space, known as coronal mass ejections (CMEs), are also frequently observed in association with solar flares. Although the causal role of solar flares in producing CMEs is not clear, there is an increasing association between flares and fast, wide CMEs, starting from approximately 1:3 for low C–class events and progressing to a nearly one–to–one association for flares stronger than X2.0 (Georgoulis et al., 2024). The strong shocks driven by the fastest CMEs are also effective in accelerating a small fraction of the particles they intercept to very high energies. Although only a small fraction of the solar–wind particles encountered by these shocks are accelerated, the flux of these particles relative to the cosmic–ray background can be quite high, and the accelerated particles may be detected on all field lines intersecting the shocks.

Flares occur in magnetically complex regions, known as active regions (ARs), within which sunspots represent the most conspicuous photospheric manifestation. Sunspots host strong magnetic fields and appear dark because magnetic fields inhibit convective energy transport. They are commonly organized in bipolar groups within larger ARs. The magnetic structure of ARs is not confined to the photosphere, but extends upward into the chromosphere and corona, where it shapes the overlying atmospheric configuration and plays a central role in flare–related energy storage and release. The magnetic and morphological complexity of these regions is closely tied to flare productivity, making ARs and sunspot groups the natural starting point for any observational discussion of flaring activity.

This chapter is organized as follows. Section 2.1 introduces the observational classification of sunspots and ARs and their relation to flare productivity, while Section 2.2 focuses on the formation and evolution of δ–spot configurations. Section 2.3 then discusses photospheric signatures of flare–productive ARs, and Section 2.4 extends the discussion to atmospheric and subsurface signatures. Finally, Section 2.5 describes the Solar Dynamics Observatory (SDO) data products used throughout this thesis.

2.1 Sunspots Classification and Flare Productivity

Sunspot numbers vary with a period of roughly 11 years, known as the solar cycle of activity (see Figure 2.1). At the beginning of a solar cycle, sunspots tend to appear at heliographic latitudes of about 30°–45°; as the cycle progresses, they appear at lower latitudes, a pattern known as *Spörer's law* (Carrington, 1858).

The first direct evidence that sunspots possess strong magnetic fields was provided by Hale (1908), who demonstrated Zeeman broadening and splitting of magnetically sensitive sunspot spectral lines, including the Fe I 6301.72 and 6302.71 Å. Hale et al. (1919) then showed that sunspot groups are typically bipolar magnetic systems whose leading and following spots usually have opposite polarities, that this polarity pattern differs between hemisphere

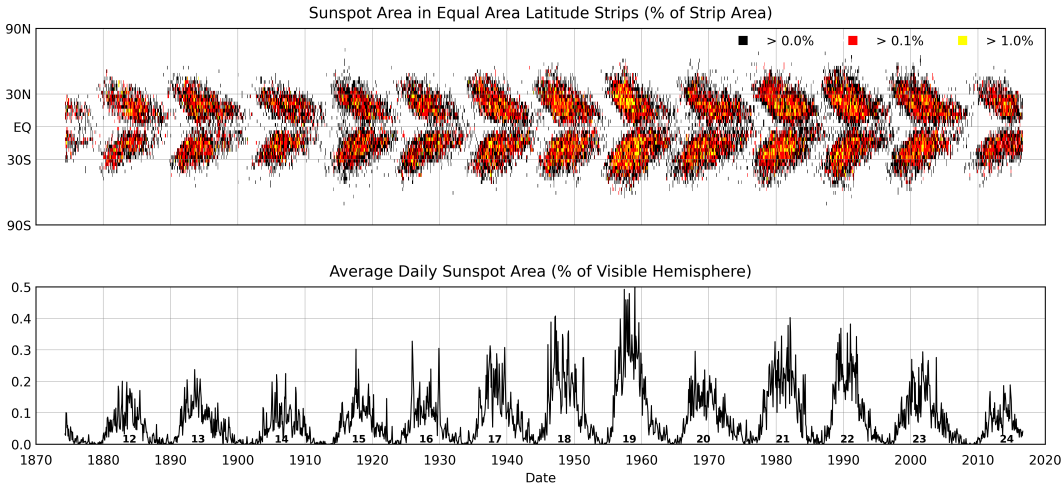


FIGURE 2.1: Temporal and latitudinal evolution of sunspot activity from 1874 to 2016. The upper panel shows the butterfly diagram of sunspot area in equal–area latitude strips, expressed as a percentage of each strip area, illustrating the systematic equatorward migration of sunspot emergence during successive solar cycles, in accordance with Spörer's law. The lower panel shows the corresponding disk–integrated average daily sunspot area, expressed as a percentage of the visible solar hemisphere, with solar cycle numbers indicated near each activity maximum.

and reverses across solar cycles, and that bipolar groups exhibit a systematic latitudinal tilt now known as *Joy's law*. The same study also introduced a magnetic classification of sunspots based on 970 groups observed during 1915–1917. These observational results were later formalized by Hale and Nicholson (1925) into the law of sunspot polarity, or *Hale–Nicholson law*.

These early observational results established sunspots not only as morphological features, but also as tracers of organized magnetic structure, naturally leading to the development of different classification schemes. The Mount Wilson magnetic classification of sunspots is a scheme based primarily on magnetic polarity measurements and supplemented by the distribution of associated calcium and hydrogen flocculi (Hale and Nicholson, 1938). Sunspot groups were divided into three principal classes, α, β, and γ. α–groups are unipolar, consisting of a single spot or a set of spots of the same magnetic polarity; these are further subdivided into α, αp, and αf according to whether the surrounding flocculi are roughly symmetric, elongated behind the group, or elongated ahead of it. β–groups are bipolar and consist of preceding and following portions of opposite polarity; they are subdivided into β, βp, and βf according to whether the two ends are roughly equal, dominated by the preceding member, or dominated by the following member. $\beta\gamma$–groups retain an overall bipolar organization, but the dividing line between the preceding and following polarities is not clearly defined, or small opposite–polarity companions complicate the pattern. γ–groups are complex, with magnetic polarities too irregularly distributed to fit the bipolar

pattern.

One of the earliest systematic statistical studies of the relationship between solar eruptions and sunspot groups, based on observations from 1935–1937, was presented by Giovanelli (1939). He defined the eruption probability as the average number of eruptions observed per day per sunspot group and showed that this probability increases approximately linearly with sunspot–group area. He also found that eruption productivity increases with the magnetic complexity of the sunspot–group type, being lowest for αp groups, approximately the same for βf, β, and βp groups, higher for $\beta\gamma$ groups, and highest for γ groups. The comparable activity of βf and βp groups suggested that the orientation of the bipolar group has little influence on its activity, whereas increasing magnetic complexity has a stronger effect. These findings were supported by subsequent studies (Bell and Glazer, 1959; Greatrix, 1963)

The Mount Wilson classification was later modified by the introduction of the δ designation for sunspot groups containing umbrae of opposite magnetic polarity within a common penumbra (Künzel, 1959; Künzel, 1965). The work of Künzel (1959) established a clear statistical connection between magnetic complexity and flare productivity, and argued that local magnetic–field structure is more directly related to flare occurrence than spot–group morphology alone. Later statistical studies showed consistent results (e.g. Sammis, Tang, and Zirin (2000), Ternullo et al. (2006), Toriumi et al. (2017), and Chaudhari et al. (2024)). Other studies have also shown correlations between δ–spots and flares associated with energetic proton emission (Warwick, 1966; Sakurai, 1970), γ–ray emission (Xu et al., 1991), and fast CMEs (Wang and Zhang, 2008).

The Mount Wilson classification scheme is based on the magnetic structure of sunspots, but early attempts to classify sunspot groups were based mainly on their morphology as observed in white light, which for many years provided the primary basis for describing sunspot structure and evolution. The Zurich classification is a morphological scheme developed by Waldmeier (1938) from the earlier classification introduced by Cortie (1901). This approach was later refined in the McIntosh classification (McIntosh, 1990). The McIntosh classification has the form Zpc, where Z is a modified Zurich class describing the evolutionary stage and large–scale structure of the sunspot group, p characterizes the penumbra and symmetry of the largest spot, and c describes the compactness of the spotted interior between the leading and following parts of the group. By combining these three components, McIntosh expanded the traditional Zurich scheme into 60 distinct sunspot–group types, with more objective definitions based solely on white–light observations.

A major contribution of the paper McIntosh (1990) is that the added descriptors capture features of sunspot groups that are closely related to size, maturity, stability, and complexity, and therefore correlate better with solar flares than the older Zurich classification alone . Using statistics from 1969–1976, McIntosh showed that flares are concentrated in the larger and more complex types, especially groups with large asymmetric principal spots and

compact interiors, while symmetric and open groups are comparatively inactive. Bornmann and Shaw (1994) showed that flare rates increase along the main diagonal of the three–dimensional McIntosh classification space, from the simplest classes, such as A/B/Hxx, to the most complex classes, such as Fkc. Other studies have also reported that morphological complexity correlates with flare occurrence (Gallagher, Moon, and Wang, 2002; Lee et al., 2012; McCloskey, Gallagher, and Bloomfield, 2016; Janssens et al., 2025). The McIntosh classification system is currently used by the Space Weather Prediction Center of the National Oceanic and Atmospheric Administration (NOAA) (Leka et al., 2019; Camporeale and Berger, 2025).

2.2 Formation and development of δ–spots

Evidence of the role of active–region magnetic evolution in flare productivity was provided by Martres et al. (1968), who showed that flare onset is typically associated with neighboring opposite–polarity magnetic structures within ARs, and that these flare–producing pairs characteristically evolve in opposite senses, with one element growing while the other declines. Using 18 years of observations from Big Bear Solar Observatory (BBSO), Zirin and Liggett (1987) analyzed 21 δ sunspot groups and argued that great flares are fundamentally tied to δ configurations, especially those in which substantial opposite–polarity umbrae are forced into close contact within a common penumbra. They classified δ–spots into three distinct classes according to how they formed.

The first type consists of a complex of spots emerging simultaneously as intertwined dipoles, referred to as "island δ" configurations (e.g. McMath 11976 Zirin and Tanaka (1973) and NOAA 5395 Wang et al. (1991)). The second type forms in two stages when large satellite dipoles emerge within or near the penumbra of an older, dominant spot, producing a δ configuration (e.g. Mount Wilson 21030 Patterson and Zirin (1981) and NOAA 10930 Kubo et al. (2007)). The third type results from the collision between two separate growing bipolar groups (e.g. NOAA 11158 Toriumi et al. (2014) and NOAA 10314 Poisson et al. (2013)). Zirin and Liggett (1987) also suggested that δ configurations are not merely surface arrangements of spots, but manifestations of a deeper subsurface process.

This interpretation was further developed by Tanaka (1991) through the study of the isolated δ group McMath 13043, observed in July 1974, using evolutionary observations from BBSO. The group exhibited unusually rapid evolution accompanied by numerous intense flares, providing an exceptional case for examining the relation between flare occurrence and changes in magnetic configuration. Its unusual surface evolution was interpreted as the photospheric expression of a complex subsurface twisted magnetic–rope system, with the observed sequence of spot and magnetic–structure changes reflecting the continued emergence of a complexly intertwined magnetic flux rope. In this framework, successive changes in spot morphology and magnetic connectivity correspond to the photospheric intersections of different parts of an emerging three–dimensional twisted rope or knot. The concentrated

flare activity of isolated δ groups was therefore attributed to internal magnetic activity below the surface, capable of generating anomalous twisted ropes whose emergence drives rapid spot motions, extreme shear, and repeated major flares.

Poisson et al. (2013) studied the emergence and evolution of the highly flare–productive NOAA 10314 and concluded that its complex quadrupolar structure and sustained activity were produced by the emergence of a single, severely deformed magnetic flux tube rather than by two unrelated bipoles. The morphology of the flux tube was inferred from the photospheric evolution of the magnetic polarities observed in Solar and Heliospheric Observatory (SOHO)/Michelson Doppler Imager (MDI) magnetograms. The region began with the appearance of two closely timed bipoles, whose inner polarities packed together to form a persistent δ–spot. NOAA 10314 produced 37 flares, including two X–class and seven M–class events. During the major flares, a complete restructuring of the coronal configuration was observed. In particular, before the X1.5 flare on March 18, 2003, the coronal structure was clearly sigmoidal, with pronounced shear in the region coincident with the inversion line of the δ–spot. After the flare, the coronal loop configuration became more potential, indicating a relaxation of magnetic stress.

A method for reconstructing the subsurface three–dimensional magnetic structure of an emerging AR directly from observations was presented by Chintzoglou and Zhang (2013), using high–cadence SDO/Helioseismic and Magnetic Imager (HMI) magnetograms of NOAA 11158. By applying an image–stacking technique to the time sequence of magnetograms, they generated a three–dimensional data cube of the emerging flux system. They found that, despite its complex photospheric appearance, the AR consisted of two major bipoles comprising four main polarities, each of which displayed a tree–like morphology, with a relatively coherent, trunk–like root at depth that branches into multiple thinner flux tubes near the surface. The roots of the four main polarities were approximately aligned along a straight line, whereas the upper branches were somewhat non–coplanar, indicating that much of the apparent surface complexity develops in the upper layers during flux emergence. The study therefore concluded that an AR with a highly complex photospheric morphology may nevertheless originate from a comparatively simple, straight subsurface flux tube that undergoes both horizontal and vertical bifurcation during its ascent through the convection zone.

A broader survey by Toriumi et al. (2017) examined all flares of class $\geq$ M5 occurring within 45° of disk center over six years from May 2010 and classified the host ARs into four groups according to their development. The first three were equivalent to the evolutionary types introduced in Martres et al. (1968). However, some X–class events were found to occur between two independent ARs and therefore did not require the presence of a δ sunspot or a strong–gradient polarity inversion line (PIL). These cases were categorized in a fourth group, termed inter–AR events. For example, the X1.2 event of 2014 January 7 occurred in the magnetic environment of neighboring NOAA 11944 and 11943, with the source region located near their mutual boundary and assigned to NOAA 11944 (Möstl et al., 2015; Wang et al., 2015).

More recent studies have complemented these earlier case–based interpretations with broader statistical evidence. Moore et al. (2026) investigated the formation of δ sunspots with sharp PILs using more than a decade of SDO/HMI full–disk magnetograms and continuum images, and found that only one of the 28 sharp–PIL δ–sunspots in their sample formed from a single newly emerged bipolar magnetic region, whereas the remainder resulted from the interaction and merger of two or more emerging or already emerged bipolar magnetic regions. This result argues against the view that most sharp–PIL δ–sunspots are produced by a single writhed or kinked flux rope. Instead, the authors argue that separate magnetic flux systems emerge and are subsequently brought together by photospheric advection, until they are compressed in convective downflow, where converging motions pack opposite polarities closely enough to produce a sharp PIL and a δ configuration.

A complementary cycle–scale statistical analysis of Solar Cycles 23–25 by Louis (2025) confirmed that δ ARs are preferentially associated with M– and X–class flares. However, the study also showed that their occurrence, lifetime, and flare productivity vary substantially from cycle to cycle, and that a significant fraction remain flare–quiet. This result underscores that the δ configuration is a strong, but not sufficient, indicator of eruptive activity.

2.3 Photospheric features

2.3.1 Strong–field, strong–gradient, highly–sheared PILs and magnetic channels

Flares are generally understood to result from the sudden release of magnetic energy through magnetic reconnection. Such a process is favored where magnetic fields of opposite polarity are brought into close contact, as in PILs; it is therefore natural that flares are commonly observed to occur around them. Strong–field, strong–gradient, and highly sheared PILs have been consistently associated with major flare activity, as they serve as proxies for compact, current–carrying magnetic structures emerging through the photosphere (Zirin and Tanaka, 1973; Hagyard et al., 1984; Wang et al., 1996; Schrijver, 2007).

Both the vertical and transverse magnetic fields around strong PILs can reach very large values. Direct spectropolarimetric measurements by Zirin and Wang (1993b) showed that strong transverse fields in δ–spots are generally aligned parallel to the neutral line and can reach values as high as 3980 G, comparable to umbral fields. Later observations confirmed that such intense fields can occur within or near light bridges between opposite–polarity umbrae. Jaeggli (2016) reported a persistent, nearly horizontal field of about 3500–3800 G near the PIL of NOAA 11035, while Wang et al. (2018a) measured a transverse field of about 5570 G in the light–bridge structure of NOAA 12673. Earlier and more recent surveys further show that fields exceeding 4000 G are rare but are often associated with complex sunspot groups and light bridges; one example is the reported super–strong magnetic field of

6250 G in NOAA 11967, one of the strongest photospheric magnetic fields directly observed on the Sun (Livingston et al., 2006; Okamoto and Sakurai, 2018).

The vertical magnetic field changes sharply across the PIL, indicating that opposite–polarity flux systems are brought into close proximity and that magnetic flux is concentrated within a compact, strongly stressed boundary layer (Moreton and Severny, 1968; Wang et al., 1991; Wang, Xu, and Zhang, 1994). In flare–productive ARs, magnetic gradients across the PIL can reach values of the order of several hundred G Mm^{-1} (Liu et al., 1996; Wang and Li, 1998; Cui et al., 2006; Jing et al., 2006; Song et al., 2009).

The shear angle of the vector magnetic field is defined as the angle between the observed magnetic field and its corresponding current–free potential field, where the potential field satisfies the boundary conditions provided by the observed line–of–sight (LoS) field (Hagyard et al., 1984; Lü, Wang, and Wang, 1993). It is therefore a measure of the non–potentiality of the magnetic configuration and, by extension, a proxy for the electric currents and magnetic free energy stored in the AR. In a potential, non–sheared field configuration in an AR, the transverse field direction in the vicinity of the neutral line is expected to be approximately normal to the line, compatible with loops arching across the neutral line from positive to negative polarities. However, shear angles of 80°–90° are frequently observed around PILs during flares (Hagyard, Venkatakrishnan, and Smith Jr, 1990; Hagyard, 1990).

Another important feature of flaring PILs is the so–called magnetic channel, defined as an elongated magnetic structure with alternating polarities. Strong transverse magnetic fields are also found along these channels. Surface plasma flows are observed along the channels as well, and the structure is associated with new flux emergence (Zirin and Wang, 1993a; Wang et al., 2002). The spatial resolution of early magnetograms was insufficient to study magnetic–channel structure in detail, but the launch of the Hinode spacecraft enabled such investigations. In particular, the Solar Optical Telescope (SOT), on board Hinode, provided unprecedented observations with high spatial and temporal resolution, allowing more advanced studies of the nature of magnetic channels and their role in powering flares (Wang et al., 2008; Lim et al., 2010).

Small–scale magnetic disturbances embedded in strongly sheared PIL regions have also been shown to trigger major flares (Bamba, Yamamoto, and Okamoto, 2013). The role of such disturbances in flare triggering was later investigated using SDO/HMI and SDO/Atmospheric Imaging Assembly (AIA) observations in both detailed event studies and broader assessments (Toriumi et al., 2013; Bamba et al., 2017; Bamba and Kusano, 2018).

2.3.2 Photospheric Magnetic Complexity and Intermittency

It is well established that the frequency and magnitude of solar flares increase with the size and with the morphological and magnetic complexity of their host ARs, as reflected by the McIntosh white–light and the Mount Wilson magnetic classifications. Over the past decades, the characterization of

solar magnetic fields has evolved from morphological descriptions toward a quantitative view in which photospheric magnetic fields are understood as multifractal, intermittent, and scale–dependent structures.

Lawrence, Ruzmaikin, and Cadavid (1993) showed that photospheric magnetic fields are better described as multifractal measures than by a single fractal dimension. Quiet–Sun fields appear partly Gaussian but also strongly intermittent, suggesting that they are generated by small–scale convective dynamo action, while active–region fields show broader multifractal spectra that reflect the contribution of strong, sunspot–like fields together with weaker large–scale magnetic components. The study therefore frames solar magnetic structure as scale–dependent and intermittent, rather than statistically homogeneous.

The multifractal interpretation of solar magnetic fields was subsequently extended by Cadavid et al. (1994). By comparing observed magnetograms with analytical and numerical models, they showed that observed LoS magnetic–field distributions can be interpreted as a combination of a self–similar multifractal component and a Gaussian white–noise–like contribution. A more rigorous framework for describing the scale–dependent and intermittent structure of photospheric magnetic fields was later provided by Lawrence, Cadavid, and Ruzmaikin (1996), who showed that different measures, such as generalized dimensions, box–counting dimensions, and threshold dimensions, capture different aspects of the magnetic–field distribution and should not be treated as equivalent.

Building on earlier work that established the fractal and multifractal nature of solar magnetic fields, McAteer, Gallagher, and Ireland (2005) shifted the emphasis from structural characterization to predictive relevance by relating active–region fractal dimension to flare occurrence and magnitude. Using active–region subfields extracted from SOHO/MDI full–disk magnetograms, they applied an automated box–counting method to 9342 active–region images spanning 2742 days and compared the resulting fractal dimensions with both Mount Wilson magnetic classifications and flare productivity. They found that ARs with larger fractal dimensions tend to be more flare–productive. At the same time, the fractal–dimension distributions of different Mount Wilson classes α, β, $\beta\gamma$, $\beta\gamma\delta$ were found to be broadly similar, implying that a single global fractal dimension captures only part of the magnetic complexity relevant to flaring.

Photospheric magnetic complexity can also be interpreted in the language of turbulence theory, which reformulates the earlier multifractal description in terms of intermittency and magnetic–energy dissipation. In this framework, Abramenko et al. (2002) showed that the scaling of the vertical photospheric field is not universal across ARs, but instead reflects different levels of intermittency that correlate with flare productivity. Their analysis suggests that the relative contribution of small–scale fluctuations in magnetic–energy dissipation becomes larger as an AR evolves toward stronger flaring states, indicating that flare–related magnetic reorganization involves not only large–scale structures but also the progressive buildup of small–scale dissipative

structures before flare onset. Abramenko et al. (2003) computed two turbulence parameters from time sequences of photospheric magnetograms during major flares, consisting of the scaling exponent β, derived from sixth–order structure functions and used as a measure of magnetic intermittency, and the correlation length λ of the magnetic–energy–dissipation field $\epsilon(B_z)$. Their results suggested the presence of a preflare turbulent phase and indicate that the peak in λ may trace an avalanche–like sequence of coronal reconnection events within the framework of self–organized criticality.

Structural complexity in ARs can also be characterized through the magnetic power spectrum. Using high–resolution SOHO/MDI magnetograms of 16 ARs, Abramenko (2005) computed the power–law index α of the magnetic power spectrum, $E(k) \propto k^{-\alpha}$, where k denotes the spatial wavenumber, and compared it with the daily soft X–ray flare index A. The flare index is a weighted measure of flare productivity based on GOES B–, C–, M–, and X–class events, defined as

$$A = \frac{1}{\tau}\left[100\sum_i I_{X,i} + 10\sum_j I_{M,j} + 1.0\sum_k I_{C,k} + 0.1\sum_l I_{B,l}\right], \qquad (2.1)$$

where τ is the observing interval. Abramenko (2005) showed that flare–productive regions have systematically steeper spectra; in particular, ARs producing X–class flares typically have $\alpha > 2$, whereas flare–quiet regions are closer to the Kolmogorov value $\alpha \approx 5/3$. In addition, the spectral slope measured early in the emergence phase was found to be similar to its later average value, indicating that the magnetic power spectrum may contain information about the future flare productivity of an AR from an early stage. Further evidence for the connection between intermittency and flare productivity was provided by Abramenko and Yurchyshyn (2010) through the analysis of a large sample of 214 ARs.

From a practical point of view, however, Georgoulis (2005) concluded that simple fractal measures and generalized multifractal correlation dimensions of AR magnetic fields are of limited value for flare prediction, whereas the structure functions $S_q(r)$ and, in particular, the temporal evolution of their inertial–range scaling exponents $\zeta(q)$, may provide statistically useful diagnostics of imminent flaring activity. Overall, a substantial body of work has examined the fractal and multifractal properties of solar active–region magnetic fields (see the review by McAteer, Gallagher, and Conlon, 2010).

2.3.3 Current Neutralization and Multiscale Current Structure

Magnetic energy released in solar flares and CMEs is stored in the non–potential field of ARs, and a long–standing question is whether the associated electric currents are neutralized within each magnetic polarity or whether a net current survives on active–region scales. In a series of influential papers, Melrose (1991), Melrose (1995), and Melrose (1996) argued, on the basis

of vector–magnetogram interpretations, that flare–productive regions commonly exhibit unneutralized currents, with current flowing from one footpoint to the other without an observed coronal return current. He further maintained that this poses difficulties for standard flare models in which post–emergence twisting or shearing of isolated flux tubes should produce locally neutralized current systems and that explicit consideration of current paths and boundary conditions is therefore essential for understanding energy storage and release. By contrast, Parker (1996) argued that flare–related evolution should be formulated primarily within the standard magnetohydrodynamic (MHD) description in terms of the magnetic field and plasma velocity, with current treated as a derived quantity. He also questioned whether magnetogram–based inferences of net current are physically reliable, given the unresolved fibril structure of the photospheric field. Melrose's reply clarified that the principal disagreement was not over the validity of MHD itself, but over the physical interpretation of vector–magnetogram data, the usefulness of the "current viewpoint" in constraining boundary conditions, and the implications of apparent non–neutralization for flare models. Together, these papers established the current–neutralization problem as a fundamental issue in solar physics, linking the origin and closure of active–region currents to the buildup of free magnetic energy and to the initiation of eruptive activity.

Two basic mechanisms have been proposed for the development of non–neutralized currents in solar ARs, namely the emergence of current–carrying magnetic flux and the buildup of such currents by typical photospheric flows, such as twisting and shearing motions. Leka et al. (1996) concluded from the morphology, proper motions, and vector–magnetogram measurements of emerging bipoles in NOAA 7260 that the newly emerging flux was already carrying substantial electric current before it became visible at the solar surface. Their results therefore support the view that, at least in some ARs, electric currents and magnetic free energy are brought to the photosphere by the emergence of pre–stressed, current–carrying flux, rather than being generated only afterward by photospheric motions. Dalmasse et al. (2015) used three–dimensional MHD simulations to show that typical photospheric twisting and shearing motions generate both direct and return currents, but that a net coronal current can nevertheless develop when magnetic shear builds up along the PIL. Their results imply that complete current neutralization is a special case, whereas non–neutralized currents are expected in more realistic active–region configurations with a sheared PIL. Observational evidence for near–neutralized photospheric currents in sunspots has been reported, although it remains limited (Wilkinson, Emslie, and Gary, 1992; Venkatakrishnan and Tiwari, 2009).

Subsequent work revisited the current–neutralization problem using higher–quality vector magnetograms and modern three–dimensional modeling, increasingly pointed to a close connection between net currents, magnetic shear, and eruptive activity. Georgoulis, Titov, and Mikić (2012) reported that non–neutralized current patterns are preferentially concentrated near strong PILs and argued that these currents are physically linked to the Lorentz force

responsible for the generation of magnetic shear, thereby supporting the view that net currents are intrinsic to strongly interacting magnetic polarities rather than mere observational artifacts. Extending this observational line, Kontogiannis et al. (2017) found that flare–productive ARs tend to exhibit stronger non–neutralized currents, suggesting that departures from current neutralization are not only common in complex ARs but are also relevant to their flare potential. A similar conclusion was reached by Liu et al. (2017b), who showed that electric–current neutralization, magnetic shear along the PIL, and eruptive behavior are closely coupled, with more weakly neutralized current systems tending to be associated with more eruption–prone magnetic configurations.

A recent follow–up that revisits this long–standing question of current neutralization in newly emerging solar ARs using a much larger sample and more advanced vector–magnetogram data was provided by Duan, Xing, and Jiang (2024). Using SDO/HMI observations of 86 well–isolated flux–emerging ARs, they found that most ARs are born with currents close to neutralization, while the strongest departures from neutrality occur in the more non–potential regions, thereby linking departures from current neutralization primarily to magnetic non–potentiality rather than to active–region size or flux–emergence rate. At the same time, Parker's reservations regarding the interpretation of magnetogram–derived net currents have continued to inform later work. Sun and Cheung (2021) showed that some apparently non–neutralized photospheric currents may result from a geometric projection effect during flux emergence. Their work does not exclude genuine net currents, but it cautions that magnetogram–based current imbalance is sensitive to the geometry and stage of emergence and should not always be interpreted as direct evidence of physically non–neutralized currents.

Another line of investigation has addressed the role of current helicity in the characterization of active–region magnetic structure. In its general local form, current helicity is defined as

$$H_c \propto \mathbf{B} \cdot (\nabla \times \mathbf{B}), \tag{2.2}$$

and therefore quantifies the degree of alignment between the magnetic field and the electric current density. In photospheric applications, where only the vertical component of the current density can be inferred directly from the horizontal derivatives of vector magnetograms, current helicity is commonly represented by the proxy

$$h_c \propto B_z (\nabla \times \mathbf{B})_z, \tag{2.3}$$

In contrast to unsigned current, current helicity conveys not only the presence of electric currents but also their sign, thereby providing information on the dominant chirality or sense of twist of the magnetic field. It is thus regarded as a diagnostic of magnetic non–potentiality and of the free magnetic energy stored in ARs. Moreover, the spatial distribution of current helicity provides insight into the organization of oppositely signed current systems and the extent to which they cancel across spatial scales.

Studies based on current–helicity measurements and cancellation analysis have consistently indicated that flare–productive ARs undergo a systematic multiscale reorganization of the photospheric current–helicity distribution. This reorganization is manifested as measurable changes in the degree of sign cancellation across spatial scales and has been interpreted as evidence for modifications in the topology, coherence, and dissipative structuring of current–carrying magnetic fields as flaring activity intensifies (Abramenko, Wang, and Yurchishin, 1996; Abramenko, Yurchishin, and Carbone, 1998; Yurchyshyn, Abramenko, and Carbone, 2000).

Sorriso-Valvo et al. (2015) represented a significant methodological advance in this line of work by exploiting the high spatial resolution, uninterrupted temporal coverage, and uniform cadence of SDO/HMI vector magnetograms to track the multiscale evolution of current density and current helicity in NOAA 11158 from flux emergence through major flaring activity. Their analysis enabled, for the first time, a detailed time–dependent characterization of cancellation properties over several days, including the first clear detection of sign singularity not only in current helicity but also directly in the vertical current density, thereby providing a basis for more detailed studies of the short–timescale relationship between photospheric magnetic complexity and major flares.

2.4 Atmospheric and subsurface evolutions

2.4.1 Sigmoids and Filaments as Observational Signatures of Flux–Rope Topology

Non–potential coronal magnetic topologies, characterized by shear, twist, and substantial field–aligned currents, can store magnetic energy sufficient to power moderate–to–large flares. When observed in soft X–rays, such coronal fields often display forward or inverse S–shaped structures known as sigmoids, which are generally interpreted as manifestations of magnetic flux–rope topology.

One of the earliest explicit observational arguments linking sigmoid brightenings to kinked magnetic flux ropes was provided by Rust and Kumar (1996). Using Yohkoh soft X–ray observations, they catalogued 103 transient brightenings near disk center, many of which exhibited clear S or reverse–S morphology, and showed that the aspect ratios of 49 well–defined events were consistent with a simple model of a twisted flux rope undergoing helical kink instability, with the observed values of $2\pi R/L_{obs}$ peaking near the value expected from the model once projection effects were taken into account. On this basis, they argued that transient sigmoids constitute observational evidence for unstable, kinked magnetic fields and suggested that the same instability may play a causal role in solar eruptions and CMEs. They further showed that reverse–S brightenings were strongly favored in the northern hemisphere, whereas forward–S brightenings dominated in the southern hemisphere, and interpreted this hemispheric pattern as evidence

that the coronal fields associated with transient sigmoids are systematically twisted. The inferred sign of helicity is consistent with the hemispheric helicity pattern previously identified in studies of H_α filaments and sunspot regions.

The statistical demonstration that sigmoidal ARs are more eruption–prone than non–sigmoidal ones, with sigmoid morphology and sunspot area acting as independent predictors of eruptive activity, established sigmoids as practically useful coronal precursors of solar eruptions (Canfield, Hudson, and McKenzie, 1999). This statistical link between sigmoid morphology and eruptive activity helped motivate their subsequent interpretation as signatures of twisted, helicity–bearing coronal magnetic fields whose eruption may involve the formation or destabilization of magnetic flux ropes (Canfield, Hudson, and Pevtsov, 2000).

Green, Kliem, and Wallace (2011) provided a clear evolutionary example of flux–rope formation in a decaying bipolar AR, showing that sustained flux cancellation at the internal PIL over 2.5 days transformed the coronal field from a weakly sheared arcade into a highly sheared arcade and then into a continuous soft X–ray sigmoid whose middle section crossed the PIL in the inverse direction, a configuration that was interpreted as evidence for a pre–eruptive flux rope. They further argued that the amount and distribution of the cancelled flux, together with the behaviour of a hot linear feature crossing the sigmoid center, suggest that reconnection progressively built the rope and likely increased its coherence immediately before the eruption. The sigmoidal flux rope eventually erupted during a B1.4–class flare, leaving a post–eruption arcade structure in soft X–ray images, consistent with the sigmoid–to–arcade evolution reported by Sterling and Hudson (1997), Hudson et al. (1998), and Sterling et al. (2000). Comparable long–term transitions from a sheared arcade or double–J configuration to a continuous sigmoid were reported by Tripathi et al. (2009), Green and Kliem (2009), and Savcheva et al. (2012), supporting the view that, in ARs, twisted flux ropes can form in situ above the PIL through prolonged photospheric driving, particularly flux cancellation, either before eruptions or during longer–lived non–eruptive evolution.

It has been shown statistically that Hα filaments and X–ray sigmoids are commonly and closely associated spatially in ARs (Pevtsov, 2002). In a detailed case study of six eruptive events Pevtsov (2002) further showed that the sigmoid activates, erupts, and is replaced by a cusp or an arcade, whereas the associated Hα filament remains intact or undergoes only partial eruption. The principal implication is that the coronal sigmoid and chromospheric filament are closely related, but occupy different parts of the same larger magnetic configuration. This interpretation is consistent with later studies that further documented the close association between filaments and sigmoidal structures (Marqué, Lantos, and Delaboudiniere, 2002; Regnier and Amari, 2004; Rust and LaBonte, 2005).

The advent of SDO/AIA made it possible to investigate the thermal structure of flux ropes and related active–region core fields in the low corona in a

far more systematic way through differential emission measure (DEM) analysis. Thermal diagnostics based on DEM analysis have shown that erupting flux–rope cores are thermally distinct from their surroundings, with the hottest plasma concentrated in the erupting core or plasmoid, whereas the leading front and surrounding coronal structures remain comparatively cooler (Cheng et al., 2012; Hannah and Kontar, 2013). In a more methodological study, Cheung et al. (2015) introduced and validated a positive–definite DEM inversion method for SDO/AIA data and applied it to NOAA 11158, showing that SDO/AIA DEM maps can resolve the thermal organization of an AR by distinguishing high–emission–measure core loops above ~ 3.5, MK during the active flaring phase from predominantly cooler peripheral fan loops, and by tracking the strong reduction of high–temperature emission as the region decayed. Extending DEM analysis to the pre–eruptive phase, Syntelis et al. (2016) combined SDO/AIA and Hinode/Extreme–ultraviolet Imaging Spectrometer (EIS) diagnostics to study two different fast CMEs that originated during two X–class flares in NOAA 11429. They identified, in the eastern and western eruptive regions, a distinct hot component in the temperature range of $\log T = 6.8 - 7.1$ associated with the presence of flux ropes. In the eastern rope, this component increased by about an order of magnitude in the 5 h before the first CME and was accompanied by blueshifts, enhanced non–thermal widths, and densities of $4 \times 10^9 - 2 \times 10^{10} \mathrm{cm}^{-3}$, indicating gradual heating and ascent before eruption.

2.4.2 Non–thermal Broadening of EUV Spectral Lines Prior to Flares

Another possible coronal response to the long–term photospheric evolution is the pre–flare non–thermal broadening of extreme–ultraviolet (EUV) spectral lines. The measured width of an emission line contains contributions from instrumental broadening, thermal broadening, and an additional non–thermal excess component. These contributions are related by

$$W_{\mathrm{obs}}^2 = W_{\mathrm{inst}}^2 + 4 \ln 2 \left(\frac{\lambda}{c}\right)^2 \left(v_t^2 + v_{nt}^2\right) \tag{2.4}$$

where W_{obs} and W_{inst} are the observed and instrumental widths, respectively, λ is the wavelength of the emission line, c is the speed of light, v_t is the thermal velocity, and v_{nt} is the non–thermal velocity.

Soft X–ray non–thermal line broadening appears to be an early coronal manifestation of flare development, rather than merely a secondary consequence of the main hard X–ray–producing energy deposition phase (Alexander et al., 1998; Ranns et al., 2000; Harra, Matthews, and Culhane, 2001). In these studies, the broadening was observed to precede, or to begin rising prior to, the strongest hard X–ray emission, and is therefore more plausibly associated with processes operating near flare onset, such as turbulence in hot plasma near or above the flare loops. The combined evidence further suggests that the dominant source of the broadening is not located at the chromospheric footpoints, but rather in hot plasma within or above the flare

loops, with turbulence and/or evaporative upflows remaining the principal interpretations. Harra et al. (2009) examined the pre–flare coronal evolution over the three days preceding the 2006 December 13 X3.4 flare, using Hinode/EIS measurements of Fe XII 195 Å line profiles. Their main result was that coronal non–thermal line widths increased before any eruptive activity and only after the helicity injection rate had already saturated, implying that the corona undergoes measurable dynamical changes many hours before flare onset.

Imada et al. (2008) found that broad non–Gaussian EUV line profiles associated with redshifts were present in the flare arcade. They identified two principal types of broad profiles: a distorted profile produced by multiple flow components, and a symmetric profile without a resolved additional component. Their interpretation was that multiple flow components are more important in newly formed loops, particularly soon after the impulsive phase, whereas older loops tend to exhibit more symmetric broadening, possibly owing to turbulence developing where reconnection downflows interact with evaporative upflows. Although the authors did not demonstrate conclusively that turbulence is the cause, they argued that the symmetric broad profiles are plausibly associated with turbulence at the interaction site between reconnection downflows and upflowing evaporated plasma. Subsequent spatially resolved EUV studies further showed that non–thermal broadening is neither spectroscopically nor spatially homogeneous within flares, but instead reflects both unresolved evaporative footpoint flows and broader loop–top or flare–arcade dynamics (Milligan, 2011; Kawate and Imada, 2013).

2.5 The Solar Dynamics Observatory (SDO)

The Solar Dynamics Observatory (SDO; Pesnell, Thompson, and Chamberlin, 2012) is a NASA mission within the Living With a Star program, launched on 11 February 2010 into geosynchronous orbit. There are three instruments onboard: the Helioseismic and Magnetic Imager (HMI; Scherrer et al., 2012), the Atmospheric Imaging Assembly (AIA; Lemen et al., 2012), and the Extreme Ultraviolet Variability Experiment (EVE; Woods et al., 2012). The mission was designed to investigate the origin of solar variability and its impact on the heliosphere and the near–Earth environment.

SDO data products used in this study are publicly available through the Joint Science Operations Center (JSOC) [1]. The SunPy library (The SunPy Community et al., 2020) can be used to programmatically query the database and retrieve the data.

2.5.1 The Helioseismic and Magnetic Imager (HMI)

The HMI instrument provides nearly continuous observations of the full solar disk with a spatial resolution of 1.0 arcsec, sampled at 0.5 arcsec per pixel on a 4096×4096 charge–coupled device (CCD) detector. The HMI

[1] http://jsoc.stanford.edu/

instrument acquires narrow–band filtergrams at six wavelengths centered on the Fe I 6173 Å spectral line. The HMI observables are referred to as Level–1.5 data, to distinguish them from Level–0 raw images and Level–1 filtergrams, which correspond to wavelength– and polarization–specific images that have already been corrected for various effects. These observables are produced through two pipelines, called LoS and vector magnetic–field pipelines (Hoeksema et al., 2014; Couvidat et al., 2016).

The LoS pipeline uses filtergrams acquired by the HMI front camera to derive photospheric observables that include Doppler velocity, LoS magnetic field, continuum intensity, and the Fe I line width and line depth. These products are generated at a cadence of 45 s from filtergrams recorded only in right– or left–circular polarization, and are available both as definitive and near–real–time (NRT). The vector magnetic–field pipeline uses filtergrams from the side camera to compute observables from both linear and circular polarization states. Its primary observable is the set of 24 Stokes images, corresponding to the four Stokes parameters measured at six wavelengths. Filtergrams from ten 135 s sequences are combined to produce the final Stokes vector at a cadence of 720 s, so that temporal averaging improves the signal–to–noise ratio of the Stokes measurements. The standard LoS observables algorithm is also applied to the 720 s Stokes $I + V$ and $I - V$ components to produce temporally averaged LoS observables.

A major source of systematic uncertainty in HMI observables arises from the geosynchronous orbit of SDO, which produces large daily variations in the Sun–spacecraft radial velocity of ± 3 km s^{-1}. Because HMI samples the Fe I 6173 Å line at only six discrete wavelength positions, the change in the Sun–spacecraft radial velocity Doppler–shifts the solar line with respect to the instrument transmission profile over the course of the day. As a result, the measured intensities are modulated in a way that propagates into the derived Doppler velocities, LoS magnetic field, and other observables, generating spurious daily signals whose amplitude depends on disk position, field strength, and the local line profile. Although the pipeline applies velocity–dependent calibrations and corrections, these effects cannot be fully removed.

Starting from the Level–1.5 Stokes observables, the HMI vector magnetic–field pipeline derives vector magnetic–field maps by inverting, at each pixel, the wavelength–dependent Stokes profiles (I, Q, U, V) measured at six positions across the Fe I 6173 Å line. These profiles encode information on the magnetic field through the Zeeman–induced modification of the spectral–line shape and polarization state. In the operational pipeline, the inversion is performed with the Very Fast Inversion of the Stokes Vector (VFISV) (Borrero et al., 2011; Centeno et al., 2014) code. VFISV begins from a trial atmospheric model, computes the corresponding synthetic Stokes profiles, compares them with the observed profiles, and iteratively adjusts the model parameters until convergence is achieved through a Levenberg–Marquardt least–squares minimization of a χ^2 merit function. The output is not limited to a magnetic–field map, but consists of a fitted set of atmospheric parameters for each pixel, together with uncertainties and selected covariances. The

inferred parameters include the magnetic–field strength, its inclination with respect to the line of sight, its azimuth with respect to an arbitrarily chosen direction in the plane perpendicular to the line of sight, as well as kinematic and thermodynamic quantities and a geometrical parameter, the filling factor, which quantifies the fraction of the pixel occupied by a magnetic structure.

The HMI vector pipeline adopts a Milne–Eddington atmosphere model, i.e. a simplified radiative–transfer model in which the physical parameters of the atmosphere are assumed to be constant with height, except for the source function, which varies linearly with optical depth. Under these assumptions, the polarized radiative–transfer equation admits the analytical Unno–Rachkovsky solution (Borrero et al., 2011; Centeno et al., 2014; Hoeksema et al., 2014), allowing the synthetic Stokes profiles to be computed much more rapidly than in a fully stratified atmosphere. Physically, this means that the inversion does not attempt to recover the full vertical structure of the photosphere, but instead retrieves an effective atmosphere whose parameters represent average conditions over the line–formation region and best reproduce the observed line profiles. In the HMI implementation, the filling factor is fixed to unity rather than treated as a free parameter; this choice assumes that the plasma is uniformly magnetized within each pixel. This choice reflects the fact that HMI spectral sampling and spatial resolution are insufficient to disentangle filling factor and intrinsic field strength reliably in the weak–field regime; therefore, the inferred field strength should be interpreted as a pixel–averaged quantity.

After the Stokes profiles have been inverted, the magnetic–field vector is obtained, but it is not uniquely determined, because the linear polarization signals Q and U determine the transverse–field azimuth only modulo 180°. In other words, two transverse–field directions differing by 180° produce the same linear–polarization signature. This is an intrinsic ambiguity of Zeeman–based vector magnetography. To resolve this ambiguity, the HMI vector pipeline applies an additional azimuth–disambiguation step using a Minimum Energy method. Rather than treating each pixel independently, the algorithm searches for the global azimuth configuration that is most physically plausible by minimizing a quantity proportional to

$$|\nabla \cdot \mathbf{B}| + |J_z|, \tag{2.5}$$

where J_z is the vertical current density. The rationale is that a physically acceptable magnetic field should be approximately divergence–free, while spurious pixel–to–pixel azimuth reversals would generate unrealistically large electric currents. The optimization is performed using a simulated annealing algorithm, which provides a robust approach when faced with a large, discrete problem with many local minima (Hoeksema et al., 2014). Prior to disambiguation, the pipeline computes a spatially dependent noise mask, reflecting the fact that the reliability of the transverse field varies across the solar disk and with spacecraft velocity. Where the transverse field is well determined, the solution from the energy minimization is returned. In weaker–field regions, where noise dominates the retrieved field, alternative approaches

are used. In particular, the pipeline may select the azimuth most closely aligned with the potential field, assign a random disambiguation, or choose the azimuth that makes the field vector closest to radial. Consequently, the confidence of the disambiguated solution is not uniform over the full field of view.

2.5.2 The Atmospheric Imaging Assembly (AIA)

The AIA instrument provides high–resolution full–disk images of the solar atmosphere in multiple wavelength channels, extending from the transition region and corona to approximately 0.5 solar radii above the solar limb with 1.5 arcsec spatial resolution and a nominal 12 s EUV cadence. The instrument was designed to study the evolution of the magnetic environment in the Sun's atmosphere and its interaction with the embedded and surrounding plasma.

AIA consists of four normal–incidence, multilayer–coated telescopes, each equipped with a 20 cm primary mirror, an active secondary mirror, and a 4096×4096 back–thinned CCD detector. The field of view is approximately 41×41 arcmin along the detector axes, with a plate scale of 0.6 arcsec per pixel. The channels are distributed across the four telescopes in pairs: telescope 1 observes 131 Å (Fe VIII/XXI) and 335 Å (Fe XVI), telescope 2 observes 193 Å (Fe XII/XXIV) and 211 Å (Fe XIV), telescope 4 observes 94 Å (Fe XVIII) and 304 Å (He II), while telescope 3 combines the 171 Å channel (Fe IX) with the ultraviolet (UV)–visible channels at 1600 Å (C IV + continuum), 1700 Å (continuum), and 4500 Å (continuum, used primarily for co–alignment with the other telescopes). AIA provides imaging in seven EUV and three UV–visible channels, with temperature response spanning approximately 6×10^4 K to 2×10^7 K, enabling the study of plasma from the upper photosphere and transition region to hot flaring coronal conditions.

The 1700 Å channel is dominated by continuum emission from the temperature–minimum region and photosphere, while 1600 Å includes contributions from C IV line emission and UV continuum, sampling the upper photosphere and transition region. The 304 Å channel is dominated by He II and primarily images the chromosphere and transition region. The coronal EUV iron channels then map plasma over a wide thermal range: 171 Å (Fe IX) mainly samples the quiet corona and upper transition region; 193 Å (Fe XII, XXIV) samples both the corona and hot flare plasma; 211 Å (Fe XIV) and 335 Å (Fe XVI) are most sensitive to the active–region corona; 94 Å (Fe XVIII) emphasizes the flaring corona; and 131 Å (Fe VIII, XXI) has a dual response, sampling both the transition region and hot flaring coronal plasma.

The instrument was designed to combine full–disk coverage, broad thermal sensitivity, and high cadence. In its nominal observing mode, AIA acquires eight full–resolution images every 10–12 s, corresponding to one image in each routinely sampled wavelength channels. Typical exposure times range from 0.5 to 3 s, although the flight software can modify the sequence and exposure times according to observing requirements. In particular, the automatic exposure–control system reduces exposure times during flares in

order to limit detector saturation, and the observing program can be adapted to acquire data at cadences as short as 2 s over reduced fields of view for dedicated campaigns. Image stability is improved by a dedicated guide telescope and image–stabilization system on each EUV telescope, which drive piezoelectric actuators on the secondary mirror and reduce low–frequency spacecraft jitter to sub–arcsecond levels.

AIA provides high–cadence, full–disk, multi–wavelength, and multi–temperature observations that complement HMI measurements of the photospheric magnetic field and support a broad range of studies connecting the evolution of the solar atmosphere to solar activity and space weather.

2.5.3 Space–weather HMI Active Region Patches (SHARPs)

Solar ARs emerge, evolve, and sometimes disappear rapidly, while spanning a broad range of sizes and levels of complexity. To identify and track these structures in a systematic manner, an automated procedure defines HMI Active Region Patches (HARPs) from the 720 s HMI LoS magnetic–field and continuum–intensity images. HARPs provide primarily spatial information on long–lived, coherent magnetic structures at approximately the scale of solar ARs.

Each HARP is assigned a unique identifier and tracked throughout its entire lifetime or disk passage, beginning one day before the first flux emerges and continuing one day after the region has decayed. Geometrically, a HARP is represented by a rectangular bounding box in CCD coordinates, together with a bitmap specifying which pixels belong to the patch and whether they are magnetically active or quiet. The bounding box is defined so as to enclose the maximum heliographic extent reached by the region during its lifetime, thereby providing a consistent spatial reference for subsequent processing.

There is no one–to–one correspondence between HARPs and NOAA ARs. Coherent regions that are small in extent or have no associated sunspot can be detected and tracked; such faint HARPs often have no NOAA correspondence. Consequently, a single HARP may encompass more than one NOAA AR, whereas many weaker or smaller coherent magnetic structures identified by the HMI procedure have no NOAA counterpart. In addition to the geometric description, the HARP product includes basic summary properties such as total and net LoS flux, area, centroid position, and associated NOAA AR numbers when available.

HARPs are delivered in both definitive and NRT forms. Definitive HARPs are constructed only after the full temporal evolution of a region is known, so that their geometry remains consistent over the complete disk passage. By contrast, NRT HARPs are generated from the information available at a given time, and therefore their extent may change during the AR's lifetime, as separate patches may later merge. In this sense, HARPs primarily constitute a tracking and geometric framework on which subsequent higher–level HMI AR products are built.

One such higher–level product is the Space–weather HMI Active Region Patch (SHARP; Bobra et al., 2014) data series. In contrast to HARPs, which

primarily provide geometric information about the patch, SHARPs also provide cutouts of the observables and computed indices. The SHARP products are released in four main series, comprising definitive and NRT versions in both CCD coordinates and in a remapped Cylindrical Equal–Area (CEA; Sun, 2013) projection centered on the patch. Quick–look SHARP data are typically available within approximately 3 h of observation, whereas the definitive products are generated approximately five weeks later.

In the standard CCD–cutout SHARP series, the pipeline collects 31 map segments, including the principal HMI observables, namely the LoS magnetogram, Dopplergram, continuum intensity, and vector magnetogram, together with additional inversion and disambiguation quantities, uncertainty arrays, and the HARP bitmap. These maps are extracted within the HARP bounding box and therefore preserve the patch as observed in the original CCD coordinate system. In the CEA version, selected segments are further remapped from plane–of–sky CCD coordinates to a heliographic CEA projection centered on the region. In this representation, the native vector magnetic–field output of the inversion, originally expressed in terms of field strength B, inclination γ, and azimuth ψ, is transformed into the heliographic components B_r, B_θ, and B_ϕ.

In addition to the map segments, SHARPs provide a set of summary parameters designed to characterize the magnetic configuration of each AR patch at a cadence of 12 min over its full lifetime. The initial SHARP series includes sixteen indices, many of which are motivated by quantities previously associated with enhanced flare productivity, such as total unsigned flux, field gradients, vertical current density, current helicity, shear angle, and proxies for magnetic free energy (Leka and Barnes, 2003a; Leka and Barnes, 2003b; Leka and Barnes, 2007). These parameters are computed from the remapped CEA data using only pixels that are both inside the HARP and above the high–confidence disambiguation threshold, as determined from the BITMAP and CONF_DISAMBIG segments; the number of contributing pixels is recorded in the keyword CMASK. The SHARP indices are therefore intended as concise quantitative descriptors of the magnetic–field distribution and its departure from a potential–field configuration, and may be used to identify regions of interest and to support statistical studies and solar event–forecasting applications.

SHARP parameters are subject to several well–documented sources of uncertainty, primarily arising from uncertainties in the retrieval of the solar vector magnetic field. The primary vector magnetic observable provided by SDO/HMI is the Stokes vector, which must be inverted to retrieve the magnetic field. This inversion relies on simplifying assumptions about the solar atmosphere which limits the fidelity of the inferred field. In addition, the inversion algorithm tends to underestimate strong fields, leading to systematically lower magnetic field values that directly bias flux–, current–, and energy–related indices.

A further complication arises from the inherent 180° azimuth ambiguity in the transverse component. Resolving it requires a nonlocal minimization algorithm that is highly sensitive to noise in the Stokes parameters, which

vary spatially and temporally across the HMI data. Since SHARP parameters are computed only from pixels labeled as high confidence by the algorithm, the set of contributing pixels changes as a region approaches the limb, where fewer pixels exceed the noise threshold. As a result, parameter values may rise or fall simply due to center–to–limb noise variations rather than true magnetic evolution. The spatial noise pattern also varies with the spacecraft's orbital velocity, producing a clear 12 h oscillation in the number of high–confidence pixels and thus artificial fluctuations in several SHARP indices. Finally, the 12 min cadence imposes its own limitation, as temporal averaging suppresses short–timescale variability, which is especially relevant during flare evolution.

The SHARP data series are especially useful for statistical studies and solar event–forecasting applications because they provide standardized, temporally tracked maps of ARs together with quantitative measurements of their magnetic properties. By combining HMI observables, vector magnetic–field maps, and derived summary indices within a consistent patch–based framework, SHARPs make it possible to study ARs over time and across large samples, provided that the observational and processing limitations discussed above are taken into account.

Chapter 3

Fundamentals of Neural Networks

"Dip a person into one particular specialty deeply enough and long enough, and he would automatically begin to assume that specialists in all other fields were magicians, judging the depth of their wisdom by the breadth of his own ignorance. "

Isaac Asimov
The Bicentennial Man and Other Stories
(Stranger in Paradise)

Deep learning has become a central approach in modern machine learning because it enables computer systems to learn useful representations directly from data. Rather than relying entirely on hand–designed features, deep neural networks construct multiple levels of representation, in which simple patterns are combined into progressively more abstract concepts. This property is particularly important for video data, such as observations of solar active regions, where the raw input consists of high–dimensional frames ordered in time, and the relevant physical information may not be directly available from individual pixels or even individual frames, but is instead encoded in the evolution of visual features across frames.

Solar flare prediction, in particular, requires models capable of extracting informative spatial and temporal patterns from sequences of solar observations, including magnetograms, ultraviolet images, and derived physical quantities. Neural architectures provide a flexible framework for this purpose, but their successful use requires understanding how they learn, how different architectures exploit different data structures, and how they can be combined to make use of complementary sources of information.

At the same time, the use of deep models raises important methodological questions, particularly regarding interpretability. Since neural networks often contain many layers and large numbers of learned parameters, their predictions can be difficult to relate directly to physical structures in the input data. This is an important issue in scientific applications, where a model should not only provide accurate predictions but also offer insight into which regions, features, or temporal changes contribute to its decisions.

Interpretability methods therefore play an essential role in assessing whether a model has learned physically meaningful patterns rather than relying on spurious correlations or background artifacts.

This chapter introduces the fundamental neural–network concepts and architectures used as the methodological background for this thesis. Section 3.1 begins with the perceptron and the transition to multilayer networks, while Section 3.2 discusses backpropagation and gradient–based training. Sections 3.3–3.5 introduce convolutional, recurrent, and attention–based architectures. Section 3.6 discusses representation learning, while Section 3.7 examines one attribution method in detail. Finally, Section 3.8 concludes with an overview of deep–learning approaches for solar flare prediction.

3.1 From Perceptrons to Multilayer Networks

Modern neural networks "learn" by optimizing parameters (weights and biases) so that the network's outputs match desired targets (supervised learning) or satisfy a reconstruction/consistency objective (self–supervised learning, e.g., autoencoders). A useful starting point for formally understanding this learning process is the perceptron, introduced by Rosenblatt as a trainable model that adjusts connection weights to map inputs to outputs (Rosenblatt, 1958). The perceptron is the simplest setting in which "learning by updating weights" can be stated without ambiguity, because the model output depends directly on one linear decision function.

Given an input vector $\mathbf{x} \in \mathbb{R}^d$, weights $\mathbf{w} \in \mathbb{R}^d$, bias $b \in \mathbb{R}$, and a binary label $y \in \{-1, +1\}$, the perceptron computes

$$z = \mathbf{w}^\top \mathbf{x} + b, \qquad \hat{y} = \operatorname{sign}(z).$$

An example is correctly classified when the signed functional margin is positive,

$$y(\mathbf{w}^\top \mathbf{x} + b) > 0, \tag{3.1}$$

and misclassified otherwise. Rosenblatt's learning rule updates the parameters only on mistakes, pushing the decision boundary toward a configuration that classifies the current example correctly:

$$\mathbf{w} \leftarrow \mathbf{w} + \eta\, y\, \mathbf{x}, \qquad b \leftarrow b + \eta\, y, \tag{3.2}$$

where $\eta > 0$ is the learning rate. The effect of this update rule can be seen by evaluating the new score for the same training example:

$$z' = (\mathbf{w} + \eta y \mathbf{x})^\top \mathbf{x} + (b + \eta y) = z + \eta y(\|\mathbf{x}\|^2 + 1), \tag{3.3}$$

so that the signed margin increases by a strictly positive amount,

$$yz' = yz + \eta(\|\mathbf{x}\|^2 + 1). \tag{3.4}$$

Thus, every error–driven update increases the margin of the misclassified example, moving it toward (and often across) the correct side of the separating hyperplane.

However, a perceptron can only solve problems where the two classes are linearly separable. Some classification problems cannot be separated by a single straight boundary; the classic example is XOR. In XOR, the positive class consists of two diagonal corners of a square, and the negative class consists of the other two corners. No single line can separate those two sets. Consequently, a single–layer perceptron cannot represent XOR, regardless of how $\mathbf{w}$ and bias b are chosen (Minsky and Papert, 1969; Bishop, 1995).

Despite this representational limitation, which derives from the fact that the perceptron is a linear classifier, the perceptron established two enduring ideas: (i) a model can learn parameters from data, and (ii) learning can be formulated as iterative error–driven updates. Historically, this limitation motivated the move from single–layer perceptrons to multilayer perceptrons. However, the term multilayer perceptron is somewhat misleading, because such models are not simple stacks of classical perceptrons with discontinuous threshold activations; rather, they consist of affine transformations followed by continuous nonlinear activation functions, which yield richer function classes (Bishop and Nasrabadi, 2006; Goodfellow et al., 2016).

The perceptron learning rule in Equation 3.2 is effective only in the single–layer setting because the prediction depends uniquely on $\mathbf{w}$ and b, so errors provide an unambiguous update direction for these parameters. When hidden layers are introduced, the output error no longer specifies how much each intermediate unit contributed to the mistake, and the same output error can be produced by many different internal configurations. The responsibility for the error must therefore be propagated through intermediate computations. This problem is known as the credit assignment problem (Minsky, 1961), and it was recognized early as a key obstacle to training multilayer networks in a general and efficient way. This limitation constrained early neural models to architectures where learning could be expressed as local corrections or single–layer error minimization, and it contributed to skepticism about the feasibility of training general multilayer networks, thereby motivating the development of gradient–based methods that assign credit backward through the network's layers (Rumelhart, Hinton, and Williams, 1986; Goodfellow et al., 2016).

3.2 Training Deep Neural Networks

Backpropagation, short for "backward propagation of errors", is the fundamental algorithm used today for gradient–based credit assignment. It efficiently applies the chain rule to compute the derivatives of a scalar loss with respect to every parameter in a multilayer network. Gradients are computed one layer at a time, and propagated from the last layer backward, so that intermediate terms in the chain rule are reused rather than recomputed (Rumelhart, Hinton, and Williams, 1986; Goodfellow et al., 2016).

To illustrate how this algorithm operates, we will consider a simple example. In the modern neural–network literature, the perceptron introduced above can be viewed as a particular case of a more general computational unit (often called a neuron), which computes the same affine score $z = \mathbf{w}^\top \mathbf{x} + b$ but replaces the hard threshold with a (typically differentiable) activation function.

First, for each input $\mathbf{x}$, an output $\hat{y}$ is produced (forward pass):

$$z = \mathbf{w}^\top \mathbf{x} + b, \qquad \hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \quad , \tag{3.5}$$

where $\sigma(\cdot)$ is the sigmoid function. Given a target label y, we quantify how far the prediction $\hat{y}$ is from the expected output using the squared error (loss function):

$$L = \tfrac{1}{2}(\hat{y} - y)^2. \tag{3.6}$$

Then, gradients are computed by applying the chain rule along the computation graph $\mathbf{x} \to z \to \hat{y} \to L$ (backward pass, i.e., backpropagation), yielding

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial \mathbf{w}} = (\hat{y} - y)\, \hat{y}(1 - \hat{y})\, \mathbf{x}, \tag{3.7}$$

and

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial z} \frac{\partial z}{\partial b} = (\hat{y} - y)\, \hat{y}(1 - \hat{y}), \tag{3.8}$$

where we have used $\partial z / \partial \mathbf{w} = \mathbf{x}$ and $\partial z / \partial b = 1$.

Finally, the learning rule updates the parameters using these gradients with learning rate $\eta > 0$:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \, \frac{\partial L}{\partial \mathbf{w}} \quad , \qquad b \leftarrow b - \eta \, \frac{\partial L}{\partial b}. \tag{3.9}$$

Because the gradient $\nabla_\theta L$ (where $\theta = (\mathbf{w}, b)$ denotes the network parameters) points in the direction of steepest increase of the loss function in parameter space, iteratively updating θ by small steps in the opposite direction decreases L and, under suitable conditions, converges toward a (local) minimum and a set of parameters fitted to the data.

In multilayer networks, the same principle applies recursively, with gradients propagated from the output layer backward so that each layer receives an error signal that determines its local parameter updates through application of the chain rule. Intuitively, this error signal quantifies how responsible a given layer's intermediate representation is for the final loss. In the simplest case of a two–layer network with one unit per layer, the hidden unit receives credit (or blame) in proportion to the output error, modulated by how strongly the hidden representation influences the output through the subsequent–layer weights and by how locally responsive the hidden activation is to changes in its pre–activation (that is, whether it operates away from saturation). Consequently, units that weakly affect the output or operate in

regimes where the local derivative of the activation is small receive smaller updates, whereas units with a stronger influence on the output and a larger local activation derivative receive larger updates. This mechanism yields a principled distribution of credit across layers by scaling parameter updates in proportion to each layer's contribution to reducing the loss.

The basic family of gradient–based optimization methods based on this principle is known as gradient descent (GD) (Cauchy et al., 1847; Goodfellow et al., 2016). Modern deep learning relies on more efficient variants. For example, stochastic gradient descent (SGD) replaces the full–dataset gradient computation with a mini–batch estimate, greatly reducing computation per update and enabling training on large datasets; the resulting noise can also improve generalization and help escape shallow local minima (Robbins and Monro, 1951; Bottou, 2010; Goodfellow et al., 2016). Adaptive methods such as Adam (adaptive moment estimation) are extensions of SGD; they maintain running estimates of the first and second moments of the gradients, effectively using per–parameter step sizes that improve stability and speed in ill–conditioned problems and during the early phases of training (Kingma and Ba, 2014).

A crucial practical factor in training multilayer networks is the choice of activation function. Early multilayer networks often used saturating nonlinearities such as the sigmoid in Equation 3.5, which can have very small derivatives when activations saturate; as a result, parameters receive little or no update and training may converge very slowly or stall entirely. This becomes a bigger issue in deeper networks. Rectified linear units (ReLU) became popular because they mitigate saturation in the active regime and were found to improve optimization in deep networks (Nair and Hinton, 2010; Glorot, Bordes, and Bengio, 2011).

"Learning" is not only about minimizing loss, but about finding parameters that generalize to unseen data. When a deep neural network achieves low loss on the training data (the data used to fit the model) but performs poorly on unseen data, it is said to overfit, a well–known and common issue in deep learning. To mitigate this, explicit or implicit regularization is often necessary. A widely used explicit regularization method is dropout, which randomly removes units during training to reduce co–adaptation and improve generalization (Srivastava et al., 2014). Batch normalization is primarily an optimization and stabilization method that normalizes intermediate activations to speed up and stabilize training, but it can also have an implicit regularizing effect because mini–batch statistics introduce noise during training (Ioffe and Szegedy, 2015). Other commonly used regularization mechanisms include weight decay, early stopping, and data augmentation (Goodfellow et al., 2016).

3.3 Convolutional Neural Networks (CNNs)

A standard multilayer perceptron, also referred to in modern literature as a fully connected (dense) network, is a feedforward architecture composed of an input layer, one or more hidden layers, and an output layer, in which each

layer applies an affine transformation followed by a nonlinearity. The term fully connected means that every unit receives input from every component of the previous layer, so each activation mixes all features available at that stage.

For image–like data $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, a fully connected network typically requires flattening the input into a one–dimensional vector of length CHW. Flattening removes the explicit two–dimensional neighborhood structure, so the network has no inductive bias that local neighborhoods matter and no built–in notion of which pixels are adjacent on the grid. In a dense (fully connected) layer, each unit has a distinct weight for every input coordinate, so its activation is computed as a nonlinear function of a weighted sum over all input values plus a bias term. Because different input coordinates are associated with different connection weights, the same local pattern appearing at a different location, such as an edge or corner, corresponds to a different set of indices and therefore a different set of parameters. Moreover, the parameter count of the first dense layer grows with the input dimensionality, scaling as $m \times CHW$ weights, plus m bias terms, for a layer with m units, which makes dense architectures computationally expensive and more prone to overfitting on high–resolution images (Goodfellow et al., 2016).

Convolutional neural networks (CNNs) address these limitations by restricting connectivity to local receptive fields and enforcing weight sharing across spatial locations (LeCun et al., 2002). A convolutional layer computes feature maps by convolving the input with a set of learnable kernels (small arrays of learnable weights). For an input image $\mathbf{x} \in \mathbb{R}^{C_{\text{in}}\times H\times W}$, the k–th output channel at location (i, j) is

$$y_{k,i,j} = \sum_{c=1}^{C_{\text{in}}} \sum_{u,v} W_{k,c,u,v}\, x_{c,i+u,j+v} + b_k, \tag{3.10}$$

where W_k is a small kernel (e.g., 3×3 or 5×5) associated with the k-th output channel, and the same kernel weights are reused for all (i, j). This structure results in a substantial reduction in parameters relative to dense layers and makes learned features translation equivariant, meaning that translating the input produces a corresponding translation of the feature map responses. Stacking multiple convolutional layers produces hierarchical representations, in which early layers capture the simplest visual features and deeper layers capture increasingly abstract and task–relevant patterns.

Many CNN architectures incorporate spatial downsampling via pooling or strided convolutions. Reducing spatial resolution lowers computational cost and increases invariance to small translations. Together, locality, weight sharing, and multiscale feature extraction are core inductive biases underlying CNN performance in computer vision tasks (Krizhevsky, Sutskever, and Hinton, 2012; Simonyan and Zisserman, 2014; Szegedy et al., 2015; He et al., 2016; Huang et al., 2017; Ronneberger, Fischer, and Brox, 2015).

3.4 Recurrent Neural Networks (RNNs)

While CNNs exploit spatial structure, many scientific and operational problems involve sequences, such as time series or text, where the output at time t depends not only on the current input but also on the preceding inputs. Recurrent neural networks (RNNs) model such dependencies by introducing a state variable $\mathbf{h}_t$ that is updated recursively. Given a sequence $\{\mathbf{x}_t\}_{t=1}^{T}$, a standard RNN defines

$$\mathbf{h}_t = \phi(W_{xh}\mathbf{x}_t + W_{hh}\mathbf{h}_{t-1} + \mathbf{b}_h)\,, \tag{3.11}$$

where $\mathbf{h}_t$ is a hidden–state vector that summarizes information from past inputs and $\phi(\cdot)$ is a nonlinear activation (Elman, 1990; Goodfellow et al., 2016). Training proceeds by backpropagation through time (BPTT; Werbos, 2002), which applies reverse–mode differentiation to the time–unrolled recurrence to obtain gradients with respect to parameters shared across time steps.

A fundamental limitation of standard RNNs is the vanishing/exploding gradient problem. When gradients are propagated backward through many time steps, they involve products of recurrent Jacobians. If the typical Jacobian has norm smaller than 1, these products drive gradients toward zero, leading to vanishing gradients; whereas if its norm exceeds 1, they grow rapidly, leading to exploding gradients. This is a key mathematical reason why standard RNNs struggle to learn long–term dependencies (Bengio, Simard, and Frasconi, 1994). Long short–term memory (LSTM) networks address this by introducing a memory cell with gating mechanisms that regulate information flow and enable more stable gradient propagation through time (Hochreiter and Schmidhuber, 1997). In one common formulation,

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_{xi}\mathbf{x}_t + W_{hi}\mathbf{h}_{t-1} + \mathbf{b}_i),\\
\mathbf{f}_t &= \sigma(W_{xf}\mathbf{x}_t + W_{hf}\mathbf{h}_{t-1} + \mathbf{b}_f),\\
\mathbf{o}_t &= \sigma(W_{xo}\mathbf{x}_t + W_{ho}\mathbf{h}_{t-1} + \mathbf{b}_o),\\
\tilde{\mathbf{c}}_t &= \tanh(W_{xc}\mathbf{x}_t + W_{hc}\mathbf{h}_{t-1} + \mathbf{b}_c),\\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t,\\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t),
\end{aligned} \tag{3.12}$$

where $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$ are input/forget/output gates and $\odot$ denotes elementwise multiplication. The input gate $\mathbf{i}_t$ controls how much of the candidate content $\tilde{\mathbf{c}}_t$ is written into the memory cell, thereby regulating the incorporation of new information from $\mathbf{x}_t$ and $\mathbf{h}_{t-1}$. The forget gate $\mathbf{f}_t$ controls how much of the previous cell state $\mathbf{c}_{t-1}$ is retained, allowing the model to preserve long–term information when $\mathbf{f}_t \approx \mathbf{1}$ and to reset memory when $\mathbf{f}_t \approx \mathbf{0}$. The output gate $\mathbf{o}_t$ controls how much of the current cell state $\mathbf{c}_t$ is exposed as the hidden state $\mathbf{h}_t$, which is used for the next recurrent update and for downstream predictions.

By separating memory storage ($\mathbf{c}_t$) from the hidden representation ($\mathbf{h}_t$) and by enabling controlled retention and updates through $\mathbf{f}_t$ and $\mathbf{i}_t$, LSTMs provide a more stable pathway for information and gradients over long time

horizons. This characteristic made recurrent architectures, particularly LSTMs, a popular choice for sequence modeling tasks, most notably in natural language processing (Hochreiter and Schmidhuber, 1997; Sutskever, Vinyals, and Le, 2014; Bahdanau, Cho, and Bengio, 2014; Graves, Mohamed, and Hinton, 2013)

3.5 Attention Mechanisms

CNNs are designed for data with an underlying spatial organization, such as images, and build representations through local receptive fields with shared weights across spatial locations. RNNs are designed for ordered data, such as sequences, and model dependencies by propagating information through a hidden state that is updated recursively as new inputs are processed. In both cases, dependencies between distant elements are not modeled directly, but are instead mediated by a sequence of local transformations or successive state updates. Attention provides a more general mechanism for modeling dependencies by allowing the representation at one position to be updated through direct interactions with representations at other positions (Luong, Pham, and Manning, 2015; Xu et al., 2015; Vaswani et al., 2017).

An attention mechanism can be described as a function that maps a query and a set of key–value pairs to an output. The output is obtained as a weighted combination of the values, where the weights are determined by a compatibility function between the query and the corresponding keys (Vaswani et al., 2017). In this context, the query specifies what information is needed to update the current position, the keys represent the candidate positions against which that request is matched, and the values represent the information contributed by those positions once they have been judged relevant. For example, in the context of language processing, when interpreting the word *it* in the sentence "The animal did not cross the street because it was tired," the query corresponds to the information needed to resolve the referent of *it*. The other words in the sentence provide keys and values, and the attention function assigns larger weights to those considered more relevant. In this case, the position associated with *animal* would likely receive a larger attention weight, since it is the more plausible referent of *it*, whereas words such as *street* would be expected to contribute less.

Let $Q \in \mathbb{R}^{n_q \times d_k}$ denote a matrix of queries, $K \in \mathbb{R}^{n_k \times d_k}$ a matrix of keys, and $V \in \mathbb{R}^{n_k \times d_v}$ a matrix of values, where n_q and n_k are the numbers of query and key–value positions, respectively, and d_k, d_v are the key and value dimensions. In the scaled dot–product formulation introduced by Vaswani et al. (2017), the attention output is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \tag{3.13}$$

The matrix $QK^\top \in \mathbb{R}^{n_q \times n_k}$ contains pairwise similarity scores between queries and keys. Let $\mathbf{q}_i$ denote the i-th row of Q and $\mathbf{k}_j$ the j-th row of K;

then the (i, j) entry of $QK^\top$ is the dot product $\mathbf{q}_i^\top \mathbf{k}_j$. In the scaled dot–product formulation, the corresponding compatibility function is therefore

$$\alpha(\mathbf{q}_i, \mathbf{k}_j) = \frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d_k}}. \tag{3.14}$$

If $\mathbf{q}_i$ and $\mathbf{k}_j$ are closely aligned in the representation space, their dot product is large; conversely, if they are weakly aligned, the score is smaller. Applying the softmax row–wise transforms these scores into normalized weights that sum to one across all key positions for each query. Multiplication by V then yields, for each query, a weighted sum of the value vectors. Thus, the representation associated with a given query is updated by aggregating information from the entire set of candidate positions, with larger contributions from those judged more relevant by the learned compatibility scores.

The scaling factor $1/\sqrt{d_k}$ in Equation 3.13 plays an important optimization role. For large values of d_k, the dot products between queries and keys can become large in magnitude. When these large scores are passed through the softmax, the resulting distribution may become excessively peaked, so that one or a few entries dominate and the corresponding gradients become very small, making optimization difficult. Dividing the scores by $\sqrt{d_k}$ before applying the softmax stabilizes the score distribution and yields more stable training (Vaswani et al., 2017).

The core mechanism in modern attention–based architectures is self–attention, in which the queries, keys, and values are all derived from the same input representation set. If $X \in \mathbb{R}^{n \times d}$ denotes the input sequence or feature set, learned linear projections produce

$$Q = XW_Q, \qquad K = XW_K, \qquad V = XW_V, \tag{3.15}$$

where $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$ are trainable parameter matrices. Substituting these expressions into Equation 3.13 yields a representation in which each position is updated by attending to all positions, including itself, in the same input. Unlike convolution, whose interaction pattern is initially local, or recurrence, whose information flow is constrained by temporal ordering, self–attention allows direct pairwise interaction between all positions in a single layer. This makes it especially effective for capturing long–range dependencies.

One extension is multi–head attention, in which the inputs are projected multiple times using different learned parameter matrices, and several attention operations are computed in parallel. The resulting outputs are then concatenated and linearly transformed to produce the final representation (Vaswani et al., 2017). This makes it possible to attend to different relationships or features simultaneously, rather than forcing all interactions into a single compatibility structure.

Attention mechanisms can also operate between distinct representation sets. In cross–attention, the queries are derived from one source, while the keys and values are derived from another (Van Den Oord, Vinyals, et al., 2017; Jaegle et al., 2021; Alayrac et al., 2022). This allows one representation

space to retrieve task–relevant information from a different representation space. The second representation set thus provides an external context from which the first can selectively gather information. Attention therefore provides a flexible framework for relating elements either within the same set or across different sets, depending on how Q, K, and V are constructed.

In RNNs, hidden states must be computed sequentially because the state at time t depends explicitly on the state at time $t-1$. Attention, by contrast, is intrinsically parallel. The matrix products in Equation 3.13 can be evaluated simultaneously across all positions, making attention particularly suitable for modern hardware and large–scale training. This computational advantage, together with the ability to model long–range dependencies through direct pairwise interactions, explains why attention has become a central building block in modern deep learning architectures (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020; Dosovitskiy et al., 2020; Carion et al., 2020; Jaegle et al., 2021).

3.6 Representation Learning

The performance of learning algorithms depends strongly on how data are represented. In many applications, carefully designed features can make an otherwise complex task solvable with a simple model. For example, imagine being given the task of determining the time shown by an analog clock. One way to approach this would be to treat it as a computer vision problem and train a model directly on clock images to learn how to determine the time of day. However, if the clock hands are instead represented by their angles, the problem becomes a simple geometric one. Thus, in general, a good representation is one that makes the subsequent learning task easier, and the choice of representation usually depends on the specific learning task.

For complex problems involving high–dimensional data, however, it is often unclear a priori which features should be extracted. Representation learning addresses this limitation by learning intermediate variables that capture salient factors of variation directly from data, often yielding representations that transfer across tasks and reduce the need for manual feature engineering (Bengio, Courville, and Vincent, 2013; Goodfellow et al., 2016).

A canonical self–supervised approach to representation learning is the autoencoder. An autoencoder consists of an encoder $E(\cdot)$ that maps an input $\mathbf{x}$ to a latent representation $\mathbf{z}$, and a decoder $D(\cdot)$ that maps $\mathbf{z}$ back to a reconstruction $\hat{\mathbf{x}}$,

$$\mathbf{z} = E(\mathbf{x}), \qquad \hat{\mathbf{x}} = D(\mathbf{z}). \tag{3.16}$$

Training minimizes a reconstruction loss $L(\mathbf{x}, \hat{\mathbf{x}})$, so that $\hat{\mathbf{x}}$ approximates $\mathbf{x}$ under the chosen discrepancy measure.

Although an autoencoder is trained to reconstruct its input, the reconstruction $\hat{\mathbf{x}}$ is often not the only quantity of interest. Instead, the encoder

serves as a learned feature extractor, and the latent representation $\mathbf{z}$ is interpreted as a compact description of the input that retains information relevant to reconstruction while discarding nuisance variability. In this sense, the reconstruction task serves as a proxy objective that encourages $\mathbf{z}$ to encode structure sufficient to recover the input, subject to the architectural and regularization constraints imposed on the model.

If an autoencoder has sufficient capacity and is trained without additional constraints, it can approximate the identity mapping and yield a latent representation with limited informative content; therefore, useful representation learning requires restricting the solution space, for example by imposing an information bottleneck ($\dim(\mathbf{z}) \ll \dim(\mathbf{x})$), by regularizing $\mathbf{z}$, or by modifying the reconstruction objective (Vincent et al., 2008; Bengio, Courville, and Vincent, 2013; Goodfellow et al., 2016).

For image–like inputs, $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, the encoder is implemented as a stack of convolutional layers that preserve locality while progressively reducing spatial resolution, producing a latent tensor $\mathbf{z} \in \mathbb{R}^{C_z\times H'\times W'}$. The decoder then maps $\mathbf{z}$ back to the input domain using learnable upsampling operations, most commonly implemented with transposed convolutions (sometimes complemented by interpolation–based upsampling), yielding the reconstruction $\hat{\mathbf{x}} \in \mathbb{R}^{C\times H\times W}$. This setting is commonly referred to as a convolutional autoencoder (Masci et al., 2011).

Learned latent spaces also serve as computational domains for generative modeling. Vector–quantized approaches such as vector–quantized variational autoencoders (VQ–VAEs) and vector–quantized generative adversarial networks (VQGANs) learn discrete latent representations that can support scalable image generation. Latent diffusion models, in turn, enable conditional generation and controllable image–to–image transformations in the latent space of a pretrained autoencoder (Van Den Oord, Vinyals, et al., 2017; Esser, Rombach, and Ommer, 2021; Rombach et al., 2022).

In this thesis, we adopt an autoencoder–based representation–learning setting, using learned latent representations as compact descriptions of solar observations for subsequent analysis and solar flare forecasting.

3.7 Layer–wise Relevance Propagation (LRP)

Deep learning techniques have become foundational across a wide range of applications, including image recognition, natural language processing, robotics, and scientific data analysis, particularly in solar physics and space weather (Krizhevsky, Sutskever, and Hinton, 2012; Devlin et al., 2019; Levine et al., 2016; Illarionov, 2025; Camporeale, 2019). Despite their high performance, deep learning models are often regarded as opaque "black boxes" due to their lack of interpretability. This limitation is particularly problematic in domains where a simple binary output (e.g., yes/no) is insufficient. In scientific applications (such as the analysis of solar images) understanding the

conditions under which a phenomenon arises, and the mechanisms that govern its manifestation, is often more valuable than a one–dimensional, real–valued output indicating the mere presence or absence of a given feature. This opacity stems from the non–linear and hierarchical transformations that map raw input to intermediate feature representations and, ultimately, to the final classification output.

To address this interpretability gap, various attribution methods have been proposed to explain model decisions by assessing the contribution of individual inputs to the output. Examples include saliency maps (Simonyan, Vedaldi, and Zisserman, 2013), Gradient–weighted Class Activation Mapping (Grad–CAM; Selvaraju et al., 2017), Integrated Gradients (IG; Sundararajan, Taly, and Yan, 2017), Layer–wise Relevance Propagation (LRP; Bach et al., 2015), Shapley Additive exPlanations (SHAP; Lundberg and Lee, 2017). In this thesis, we adopt LRP because it provides a principled way to redistribute the prediction score back to the input domain, thereby making it possible to establish which input features most strongly support the model decision.

LRP is a backward propagation technique that decomposes the output of a model $f(\mathbf{x})$ into relevance scores R_i, assigned to individual inputs x_i. This is achieved by redistributing the prediction score backward through the network, layer by layer, using propagation rules consistent with the model's architecture, while preserving key theoretical properties such as relevance conservation, i.e., $\sum_i R_i \approx f(\mathbf{x})$, thereby ensuring consistency and interpretability (Montavon et al., 2017; Samek, Wiegand, and Müller, 2017; Montavon, Samek, and Müller, 2018). When dealing with image inputs, this process produces a pixel–level relevance map $\mathbf{R}(\mathbf{x})$ with the same spatial dimensions as the input image $\mathbf{x}$, indicating the extent to which each pixel contributes to the classification decision $f(\mathbf{x})$. Visualizing these relevance maps as heatmaps enables interpretation of which spatial regions the model considers most relevant.

Several propagation rules exist within the LRP framework. The basic propagation rule, known as the z–rule, distributes relevance in proportion to the weighted activations:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{j'} a_{j'} w_{j'k}} R_k \tag{3.17}$$

where a_j is the activation of neuron j, and w_{jk} is the weight connecting neuron j to neuron k in the next layer. To improve numerical stability when denominators approach zero, the ϵ–rule introduces a stabilizing term:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{j'} a_{j'} w_{j'k} + \epsilon \cdot \text{sign}\left(\sum_{j'} a_{j'} w_{j'k}\right)} R_k \tag{3.18}$$

where ϵ is a small positive constant that suppresses the propagation of noise in near–zero regions. A more general formulation is provided by the α–β rule, which separates positive and negative contributions:

$$R_j = \sum_k \left(\alpha \cdot \frac{(a_j w_{jk})^+}{\sum_{j'} (a_{j'} w_{j'k})^+} - \beta \cdot \frac{(a_j w_{jk})^-}{\sum_{j'} (a_{j'} w_{j'k})^-} \right) R_k \tag{3.19}$$

subject to the constraint $\alpha - \beta = 1$.

This rule allows fine–grained control over how excitatory and inhibitory signals are propagated. For example, setting $\alpha = 1$ and $\beta = 0$ considers only positive contributions, while increasing β allows for the inclusion of inhibitory effects to varying degrees. This flexibility is particularly useful in models with signed activations or weights, where negative values carry meaningful semantic information.

In the context of solar flare prediction, relevance maps can help identify which spatial regions in solar observations contribute most strongly to a model decision, allowing them to be compared with physically meaningful structures in active regions. However, these maps should be interpreted with caution, since model attributions indicate which input regions were relevant to the trained model for a specific prediction and should not be taken as direct evidence of physical causality.

3.8 Deep Learning for Solar Flare Prediction

With the increasing availability of large–scale, openly accessible solar datasets and the rapid advancement of machine learning, particularly deep learning, a growing number of studies have applied data–driven approaches to the task of solar flare prediction. To provide a conceptual overview, we group the existing methods into three broad categories, though the list compiled here is not exhaustive and these categories are not mutually exclusive.

The first category relates to feature representation strategies. Some approaches rely on manually engineered features derived from observational data, such as physical quantities computed from magnetograms (Bobra and Couvidat, 2015; Liu et al., 2017a; Nishizuka et al., 2017). Others extract features from images using fixed transformations (e.g., Gabor and random filters; Jonas et al., 2018 E. Jonas et al. 2018). A third group employs learned representations, where relevant features are automatically identified from raw data using deep learning models (Huang et al., 2018; Li et al., 2020).

The second category distinguishes methods by the type of observational data used. These include photospheric (e.g., LoS or vector magnetograms, continuum intensity), chromospheric/transition–region (e.g., Hα filtergrams and UV 1600 Å, which sample the upper photosphere and transition region Nishizuka et al., 2017; Jonas et al., 2018), and coronal (e.g., EUV channels such as 131, 171, or 193 Å, and soft X–ray imagery; see Jonas et al., 2018; Nishizuka et al., 2018; Nishizuka et al., 2020) observations.

The third category concerns the temporal structure of the input. Some models generate predictions from single observational snapshots (point–in–time prediction), while others use sequences of observations to capture the temporal evolution of active–region properties (Liu et al., 2019; Chen et al., 2019; Sun et al., 2022; Guastavino et al., 2022).

Despite the vast literature on the topic, comparing the results of different studies of deep–learning–based flare classification models remains challenging due to the absence of a standardized benchmark dataset. Each study differs in the dataset construction strategy, the types of observational data employed, the forecasting horizon, and the preprocessing methods applied to prepare the data for deep learning algorithms. These methodological differences make direct numerical comparisons inherently difficult and, in many cases, not scientifically meaningful.

A unified framework is therefore necessary to advance the field and move toward an operational setting capable of forecasting solar flares reliably and robustly, thereby helping to mitigate their potential impacts. For a more exhaustive discussion of these limitations, we refer the reader to studies that thoroughly address these issues. These include works that review or attempt to compare existing flare–forecasting methodologies (Barnes et al., 2016; Leka et al., 2019; Han et al., 2025), efforts toward constructing standardized datasets for the community (such as the multi–variate space–weather dataset introduced by Angryk et al., 2020 and the AIA Active Region Patches (AARPs) developed by Dissauer, Leka, and Wagner, 2023) and large–scale operational forecasting evaluations (Georgoulis et al., 2021), in which the diversity of methodological choices and the difficulty of establishing uniform performance baselines are illustrated. More extensive reviews of machine–learning applications in space weather, and of solar–flare forecasting more specifically, can be found in Camporeale (2019), Han et al. (2025), Han et al. (2023), and Georgoulis et al. (2024).

Chapter 4

Phase I: Representation Learning of Solar Active Regions

"The everlasting universe of things
Flows through the mind, and rolls its rapid waves "

Percy Bysshe Shelley
Mont Blanc

4.1 Introduction

The previous chapters established the observational and methodological context for this thesis. Chapter 2 discussed the observational signatures of flare–productive active regions (ARs), while Chapter 3 introduced the fundamentals of neural networks and representation learning. Building on this foundation, the present chapter begins the experimental part of the thesis by focusing on learning representations of line–of–sight (LoS) magnetograms of solar ARs that may be useful for downstream tasks, particularly solar flare prediction.

The aim of this first phase is to learn compact representations of active–region magnetic structure using a convolutional autoencoder (CAE). Rather than training the model directly for flare prediction, we first train the CAE to reconstruct the input magnetograms. In this setting, the reconstruction task provides an indirect assessment of the learned latent space, since accurate reconstruction of the magnetic morphology indicates that the latent representation retains relevant information about the spatial organization of the photospheric magnetic field.

Because magnetograms are not merely images but measurements of a physical field, reconstruction quality cannot be assessed only through pixel–level error. For this reason, the analysis in this chapter evaluates reconstruction performance at multiple levels: image–level accuracy, structural similarity, pixel–value distribution, and spectral content. In addition, we evaluate the temporal evolution of physical quantities derived from the reconstructed

magnetograms, which encode information about compact and strong flux concentrations, sharp magnetic–field gradients, and polarity inversion lines (PILs).

The chapter is organized as follows. Section 4.2 describes the data used in this study, the derived physical quantities used for evaluation, the dataset construction procedure, and the preprocessing steps. Section 4.3 presents the CAE architecture, the training objectives used in the experiments, and the evaluation tools used to assess both image reconstruction and temporal consistency. Section 4.4 reports the model–selection procedure, compares the selected training objectives on the independent test set, and examines whether the reconstructed magnetograms preserve the temporal evolution of physically meaningful magnetic–field parameters. The chapter concludes with a final section that summarizes the main conclusions of this experimental phase.

4.2 Data

4.2.1 Data Sources

The analysis presented in this chapter uses LoS magnetograms of solar ARs as the observational input. The data are obtained from the Space–weather HMI Active Region Patches (SHARP) data product, which provides active–region cutouts derived from Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) observations. Magnetograms provide a measure of the photospheric magnetic field, whose structure and evolution are closely related to flare productivity. Therefore, they constitute an excellent starting point for learning representations of active–region morphology.

To complement the image–level reconstruction analysis, we evaluate how well the reconstructions preserve physical properties of the ARs. Standard SHARP magnetic indices are generally derived from vector magnetograms, whereas the reconstruction task considered in this chapter operates exclusively on LoS magnetograms $B_{\mathrm{LoS}}(x, y)$. Consequently, quantities that require vector–field components, such as current helicity, vertical current density, magnetic shear, and free–energy proxies, cannot be computed consistently from the reconstructed magnetograms considered here. For this reason, we focus on three physical parameters that can be derived directly from $B_{\mathrm{LoS}}(x, y)$, are commonly used in the literature as flare indicators, and are closely related to SHARP indices (Bobra and Couvidat, 2015). These parameters are computed following the definitions of the corresponding SHARP indices and the reference implementation provided with the SHARPs software (Bobra et al., 2014; Bobra, Sun, and Turmon, 2021), but directly from the LoS magnetograms. They should therefore be regarded as LoS–based analogues of the corresponding SHARP diagnostics.

The first parameter is the total unsigned LoS magnetic flux,

$$\Phi = \sum_{x,y} |B_{\mathrm{LoS}}(x, y)| \, \Delta A, \tag{4.1}$$

where ΔA is the pixel area in physical units and the summation is performed over pixels belonging to the AR. This parameter measures the total unsigned magnetic flux contained in the AR and provides a global measure of the magnetic–field strength.

The second parameter is the gradient–weighted neutral line length,

$$R = \log_{10}\left(\sum_{x,y} p(x,y)\,|B_{\mathrm{LoS}}(x,y)|\right), \tag{4.2}$$

where $p(x,y)$ is a smoothed mask that weights pixels close to the inversion line. For the computation of R, the LoS magnetogram is first spatially binned by block averaging. Positive– and negative–polarity masks are then constructed by selecting strong–field pixels with $B_{\mathrm{LoS}} > 150$ G and $B_{\mathrm{LoS}} < -150$ G, respectively. The two masks are spatially smoothed and combined to identify regions where strong opposite polarities are in close proximity, thereby defining the PIL mask. This mask is subsequently smoothed with a Gaussian kernel to obtain the continuous weighting function $p(x,y)$, which assigns larger weights to pixels close to strong–field PILs and progressively smaller weights with increasing distance from them. In Equation 4.2, $B_{\mathrm{LoS}}(x,y)$ therefore denotes the rebinned magnetogram used for the computation of R. The weighted unsigned magnetic field is then summed over the valid pixels, and its base–10 logarithm gives R. This parameter measures the strength and complexity of magnetic–field concentrations near the PIL.

The third parameter is the mean LoS magnetic–field gradient,

$$\langle|\nabla B_{\mathrm{LoS}}|\rangle = \frac{1}{N}\sum_{x,y}|\nabla B_{\mathrm{LoS}}(x,y)|, \tag{4.3}$$

where N is the number of pixels belonging to the AR, the summation is performed over the same pixels, and the gradient magnitude is defined as

$$|\nabla B_{\mathrm{LoS}}| = \sqrt{\left(\frac{\partial B_{\mathrm{LoS}}}{\partial x}\right)^2 + \left(\frac{\partial B_{\mathrm{LoS}}}{\partial y}\right)^2}. \tag{4.4}$$

This parameter describes the mean horizontal variation of the LoS magnetic field over the AR and is therefore sensitive to strong local field gradients.

4.2.2 Dataset Construction

To construct our dataset, we select ARs present in the SHARP data series from May 2010 to December 2022, covering most of solar cycle 24 and the early phase of solar cycle 25. From this initial sample, we retain only SHARPs associated with exactly one National Oceanic and Atmospheric Administration (NOAA) active–region number, to enable an unambiguous cross–reference with the Geostationary Operational Environmental Satellite (GOES) flare catalog, which indexes events by NOAA active–region number.

Near the limb, increased projection effects and reduced signal quality can introduce spurious trends in both pixel–level magnetograms and integrated

indices (Bobra et al., 2014; Couvidat et al., 2016; Bobra and Couvidat, 2015). To mitigate viewing–angle–dependent systematic errors, we restrict our sample to observations of ARs within $\pm 65^{\circ}$ of the central meridian. Additionally, only good–quality records are retained, meaning that all QUALITY flags are below 65536 [1].

We query the GOES flare catalog for events in the same period. We associate each event with its corresponding SHARP record using the NOAA active–region number and select the SHARP record nearest in time to the flare peak time, within ± 12 minutes. Although the SHARP data series is sampled at a nominal cadence of 12 min, gaps may occur for several reasons, including spacecraft maneuvers, instrument calibration, ground–station handovers, eclipse seasons, and instrument anomalies or other malfunctions (see SDO/HMI coverage maps[2]). Whenever a SHARP record is missing at the time of a flare, an empty placeholder record is created to preserve the flare information. These placeholder records are not used during training because they have no associated observation, but they are retained during dataset construction to support the identification of flare–producing ARs and the subsequent split into training, validation, and test sets. This procedure yields a consistent flare–LoS–magnetogram pairing that is time–aligned with the SHARP sampling rate.

From the full population of SHARPs satisfying the observational and quality constraints described above, we identify, after flare association, the subset of ARs that produced at least one flare of any magnitude. Although this selection conditions the set of included ARs on flare activity, it does not restrict the sample to only the most energetic events. ARs are retained provided they satisfy the one–to–one NOAA mapping, central–meridian ($\pm 65^{\circ}$), and data–quality criteria, and have produced at least one flare of any magnitude. Each AR contributes many 12 min snapshots over days to weeks of evolution, the vast majority of which have no flare association even for flaring ARs. Thus, flaring ARs still include extensive quiet or low–activity intervals, including periods with relatively simple magnetic–field configurations. The final dataset is summarized in Table 4.1.

Number of Active Regions	Total Records	Weak Flares ($\leq$ C Class)	Strong Flares ($\geq$ M Class)
800	756,398	5,845 (95.07%)	303 (4.93%)

TABLE 4.1: Summary of the final dataset, including the number of retained ARs, the total number of SHARP records, and the number of weak ($\leq$ C–class) and strong ($\geq$ M–class) flare events. Percentages in the flare columns are computed relative to the total number of flares.

[1] http://jsoc.stanford.edu/doc/data/hmi/Quality_Bits/QUALITY.txt
[2] http://jsoc.stanford.edu/doc/data/hmi/coverage_maps/

4.2.3 Train/Validation/Test Split

A standard practice in machine learning is to divide a dataset into three non–overlapping subsets: a training set used to train the model, a validation set used for hyperparameter tuning, and a test set used to assess the model's performance and generalization capabilities. Ensuring that these subsets are representative of the overall data distribution is crucial for a reliable assessment of the model's performance on unseen data.

To achieve this, a common strategy is to randomly partition the dataset into the aforementioned subsets. This approach is not appropriate for ARs, as their lifespans range from hours to days. Randomly partitioning the dataset would result in samples from the same AR being distributed across the training, validation, and test sets, thereby introducing information leakage and leading to overly optimistic estimates of model performance.

One alternative is to partition the data chronologically, using earlier years for training and reserving the remaining data for validation and testing (e.g. Bobra and Couvidat (2015) and Chen et al. (2019)). However, due to the solar cycle, this strategy does not ensure a consistent distribution across subsets, as solar flares are more likely to occur during solar maximum, leading to systematic differences in event rates and active–region magnetic complexity between time periods (Wang et al., 2020).

To reduce information leakage, we partition the dataset at the active–region level, ensuring that all samples originating from the same AR belong to the same subset. This active–region–level splitting procedure is consistent with previous flare–prediction studies (Guerra, Pulkkinen, and Uritsky, 2015; Campi et al., 2019; Zheng, Li, and Wang, 2019; Li et al., 2020; Sun et al., 2022). A potential drawback of this method is the phenomenon known as sympathetic flaring, in which a flare in one AR appears to trigger flares in magnetically connected but spatially distant ARs (Pearce and Harrison, 1990; Biesecker and Thompson, 2000). Although this source of dependence cannot be avoided with any splitting strategy, the active–region–level split prevents direct information leakage by ensuring that temporally adjacent snapshots of the same AR do not appear in different subsets.

To construct the three subsets, we first separate ARs into those that produce at least one M– or X–class flare and those that do not, and randomly split each group into training, validation, and test sets using an approximate 70% / 15% / 15% target ratio. The groups are then recombined, yielding an active–region–level stratified random split that preserves a comparable fraction of M/X–flaring ARs across subsets, which helps reduce the impact of solar–cycle variability while maintaining active–region–level independence. The resulting split is summarized in Table 4.2, with the corresponding percentage distributions reported in Table 4.3.

4.2.4 Preprocessing

Our deep learning model requires consistent input dimensions; however,

Split	Number of Active Regions	Total Records	Weak Flares ($\leq$ C Class)	Strong Flares ($\geq$ M Class)
Train	559	528,270	3,978	208
Validation	144	137,677	1,125	65
Test	97	90,451	742	30

TABLE 4.2: Summary of the active–region–level train, validation, and test split, including the number of ARs, the total number of records, and the number of weak ($\leq$ C–class) and strong ($\geq$ M–class) flare events in each subset.

Split	Total Records (%)	Weak Flares (%) ($\leq$ C Class)	Strong Flares (%) ($\geq$ M Class)
Train	69.84%	95.03%	4.97%
Validation	18.20%	94.54%	5.46%
Test	11.96%	96.11%	3.89%

TABLE 4.3: Percentage distribution of the active–region–level train, validation, and test split. Percentages in the total–records column are computed relative to the full number of records, while percentages in the flare columns are computed relative to the total number of flare events within each subset.

active–region cutouts vary in size and aspect ratio, making input–size standardization necessary. Common approaches include resizing via interpolation, padding, and cropping, but each of these methods has its drawbacks (Jonas et al., 2018; Huang et al., 2018; Li et al., 2020; Deng et al., 2021; Guastavino et al., 2022; Sun et al., 2022). When resizing images to a uniform size, details can be lost, as the process involves averaging or interpolating pixel values, resulting in blurred small–scale features. Additionally, resizing can introduce geometric distortions in the magnetic structures, since the aspect ratio is not necessarily preserved. Cropping may exclude important parts of the ARs, especially since many are east–west elongated, resulting in an incomplete representation. Padding, on the other hand, can introduce artificial signals, particularly when padding values differ significantly from the actual data, potentially misleading the model.

To standardize the input dimensions while limiting the drawbacks of the approaches discussed above, we adopt a strategy based on the combined use of cropping and padding. First, we define a target size of 256×512 pixels. During training, active–region cutouts larger than this target size are randomly cropped, varying the crop location at each epoch. By doing so, the model is exposed to different subregions of the same AR, allowing it to learn from a broader spatial context without requiring explicit resizing. This procedure also serves as a form of data augmentation, effectively increasing the diversity of training samples and reducing the risk of overfitting.

For smaller active–region cutouts, padding is applied using each magnetogram's background field, which we take to be the magnetic field in regions

with no significant magnetic activity. The background magnetic field is estimated as the mean pixel value within the interquartile range (IQR). By focusing on the central 50% of the data, we reduce the impact of outliers (i.e., high magnetic–field values) and capture the central tendency of the background more accurately.

The target size is chosen as a compromise between preserving active–region structure and limiting computational cost. A smaller size would require excessive cropping of large ARs, whereas a larger size would embed small ARs within a relatively large background area and increase computational cost. The choice of a 2:1 aspect ratio reflects the typically greater longitudinal (east–west) than latitudinal extent of ARs.

Each image is scaled to $[-1, 1]$ using min–max scaling, with the global minimum and maximum estimated from the training set. This is a fundamental step in preparing the dataset, as large variations in pixel values may dominate the learning process, causing smaller but important features to be overlooked and impairing the convergence and stability of deep learning models. Estimating the scaling parameters from the training data only and keeping them fixed for the validation and test sets follows standard machine–learning practice and prevents information leakage from held–out samples. Under the usual assumption that the training, validation, and test data are drawn from the same underlying distribution, the larger and representative training subset is expected to provide a good empirical approximation of the relevant data range. Nevertheless, because the scaling extrema are estimated from the training set, values in the validation or test sets that exceed these extrema may be mapped outside the nominal $[-1, 1]$ interval. Since the same affine transformation is applied to all samples, the relative magnetic–field amplitudes across samples are preserved. This scaling strategy was preferred over image–wise normalization, which could remove physically meaningful differences in magnetic–field strength between ARs. Finally, missing values are replaced with the median of the valid pixels in each image.

4.3 Methodology

4.3.1 Model

We implement a CAE to learn compact representations of LoS magnetograms of solar ARs. Let $\mathbf{x} \in \mathbb{R}^{1\times H\times W}$ denote a single preprocessed magnetogram. The encoder $E(\cdot)$ transforms the input into a latent representation $\mathbf{z}$,

$$\mathbf{z} = E(\mathbf{x}), \tag{4.5}$$

while the decoder $D(\cdot)$ attempts to reconstruct the original input,

$$\hat{\mathbf{x}} = D(\mathbf{z}), \tag{4.6}$$

such that $\mathbf{x} \approx \hat{\mathbf{x}}$. Training is performed by minimizing a reconstruction–based objective, optionally augmented with additional regularization terms, as described in Section 4.3.2.

The CAE follows a symmetric encoder–decoder design composed of strided convolutions for downsampling and transposed convolutions for upsampling. The encoder progressively reduces the spatial resolution of the input while increasing the number of feature channels. In doing so, it aggregates local spatial information into increasingly abstract feature maps. In our implementation, the encoder is composed of five convolutional blocks. Each block contains a two–dimensional convolution followed by a rectified linear units (ReLU) activation. The convolutional layers use kernel size 3×3, stride 2, and padding 1, so that each block reduces the spatial resolution by a factor of two, resulting in an overall downsampling factor of $2^5 = 32$ (Figure 4.1). The output of the encoder is a latent tensor $\mathbf{z} \in \mathbb{R}^{512 \times H/32 \times W/32}$, which serves as the learned representation of the input magnetogram.

The latent representation $\mathbf{z}$ acts as an information bottleneck. Because it has reduced spatial resolution and lower overall dimensionality than the input, the model is encouraged to encode spatially coherent structures rather than memorizing individual pixel values. The structure and quality of $\mathbf{z}$ therefore determine not only the reconstruction fidelity but also the usefulness of the representation for downstream tasks.

The decoder mirrors the encoder using five transposed convolutional blocks. The first four blocks consist of a transposed convolution followed by a ReLU activation, while the final block uses a transposed convolution followed by a hyperbolic tangent activation. The transposed convolution layers use kernel size 4×4, stride 2, and padding 1, progressively restoring the spatial resolution of the input (Figure 4.1). The final tanh activation constrains the reconstructed output to the same normalized range used for the input magnetograms, $[-1, 1]$.

No skip connections or normalization layers are used in this model. This design choice forces the reconstruction to rely on the information encoded in the latent representation, rather than bypassing the bottleneck through direct feature transfer. As a result, differences in reconstruction quality can be more directly attributed to the training objective than to architectural shortcuts.

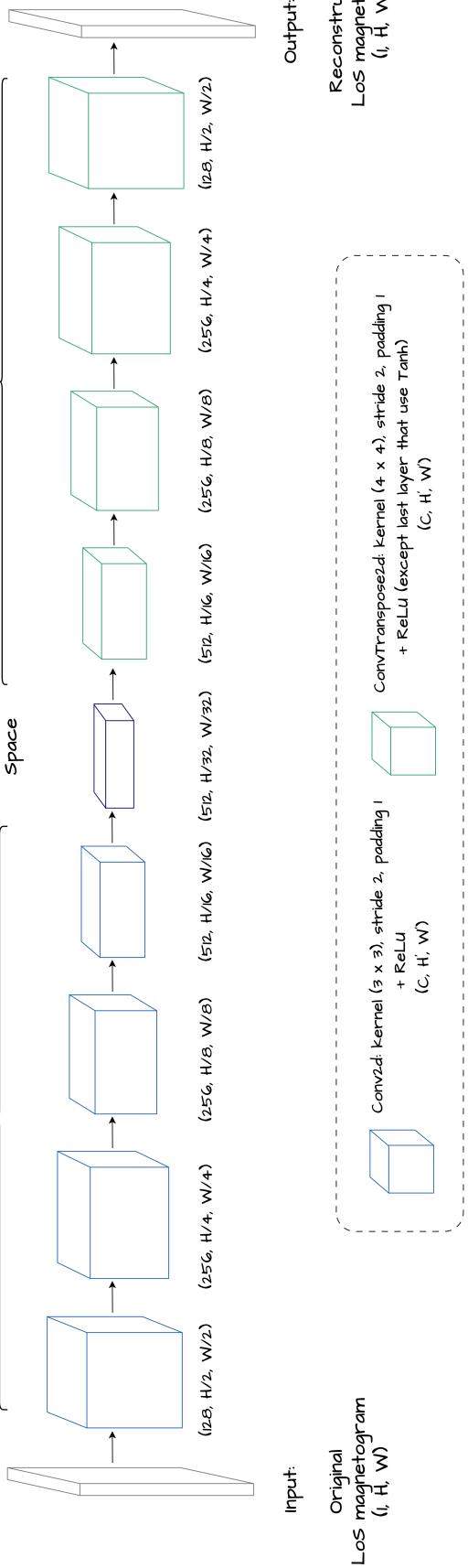


FIGURE 4.1: Schematic representation of the CAE architecture used for LoS magnetogram reconstruction. The encoder progressively reduces the spatial resolution through strided convolutions while increasing the number of feature channels. The decoder mirrors this process using transposed convolutions to reconstruct the input LoS magnetogram.

4.3.2 Loss Functions

The autoencoder is trained by minimizing a training objective that measures the discrepancy between the input LoS magnetogram and its reconstruction. Let $\mathbf{x} \in \mathbb{R}^{1\times H\times W}$ and $\hat{\mathbf{x}} \in \mathbb{R}^{1\times H\times W}$ represent the original and reconstructed magnetograms, respectively. We define the reconstruction loss, L_{rec}, to penalize pixel–wise residuals $r_{i,j} = x_{i,j} - \hat{x}_{i,j}$. For simplicity, hereafter, we let r denote a generic scalar residual such that $r \equiv r_{i,j}$ for any pixel (i, j).

The first reconstruction loss we consider is the mean squared error:

$$L_{\text{MSE}} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left(x_{i,j} - \hat{x}_{i,j}\right)^2. \tag{4.7}$$

This loss function penalizes residuals quadratically, $L_{\text{MSE}}(r) = r^2$. This formulation is optimization–friendly because the loss is smooth everywhere with a well–defined gradient $\nabla L_{\text{MSE}}(r) = 2r$. However, this quadratic growth causes large residuals to dominate both the loss and its gradient, which means that large discrepancies are corrected rapidly, but also that the training objective becomes more sensitive to outliers.

The second reconstruction loss we consider is the mean absolute error:

$$L_1 = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|x_{i,j} - \hat{x}_{i,j}\right|. \tag{4.8}$$

This loss function penalizes residuals linearly, $L_1(r) = |r|$, producing constant–magnitude gradients. Unlike L_{MSE}, L_1 is not differentiable at $r = 0$; however, in practice, this is handled by using subgradient methods in gradient–based optimization (Auslender and Teboulle, 2004). This linear growth ensures that large residuals do not receive disproportionate weight; as a result, L_1 is inherently more robust to heavy–tailed noise and sporadic artifacts.

The third reconstruction loss we consider is the Smooth L_1 loss:

$$L_{\text{SL}_1} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\phi_\beta\left(x_{i,j} - \hat{x}_{i,j}\right), \tag{4.9}$$

where ϕ_β is the Huber penalty,

$$\phi_\beta(\delta) = \begin{cases} \frac{1}{2\beta}\delta^2, & |\delta| < \beta, \\ |\delta| - \frac{\beta}{2}, & |\delta| \geq \beta. \end{cases} \tag{4.10}$$

This loss function interpolates between quadratic behavior near zero and linear behavior for large residuals, with the transition controlled by the scale parameter β (Huber, 1992). This hybrid form retains the numerical stability and smooth curvature of L_{MSE} for small residuals while limiting the influence of large residuals similarly to L_1.

Pixel–wise reconstruction losses treat each pixel residual as an independent contribution to the total error and therefore do not explicitly encode spatial organization. Consequently, reconstructions can achieve low pixel–wise

error while still exhibiting morphological deviations (Terven et al., 2025). In particular, when reconstruction is under–constrained, meaning that the model does not have sufficient information to reconstruct the input uniquely, minimizing pixel–wise reconstruction losses, especially L_{MSE}, tends to bias the prediction toward the conditional mean of the admissible solutions. This behavior can manifest visually as attenuation of sharp gradients and suppression of high–frequency detail.

In addition, low pixel–wise residuals do not guarantee consistency in the distribution of power across spatial scales. Overly smooth reconstructions may show a pronounced deficit at high spatial frequencies, indicating a systematic loss of small–scale power. Conversely, reconstructions with spurious fine–scale texture may retain comparable high–frequency power but distribute it differently, indicating a mismatch in small–scale patterning despite similar pixel–wise error.

For magnetograms, both effects can affect the recovery of relevant structures such as strong flux concentrations, sharp PILs, and fine–scale gradients, which are among the most physically meaningful features for flaring activity. To mitigate this, we augment the reconstruction loss with an auxiliary term that explicitly promotes gradient preservation in the pixel domain.

To encourage gradient preservation, we compare discrete spatial derivatives of the original and reconstructed magnetograms. The horizontal and vertical gradients of the original and reconstructed images are estimated using discrete finite differences:

$$\mathbf{x}_x = \mathbf{x} * K_x, \qquad \mathbf{x}_y = \mathbf{x} * K_y, \tag{4.11}$$

$$\hat{\mathbf{x}}_x = \hat{\mathbf{x}} * K_x, \qquad \hat{\mathbf{x}}_y = \hat{\mathbf{x}} * K_y, \tag{4.12}$$

where the finite–difference kernels are defined as

$$K_x = \begin{bmatrix} -0.5 & 0 & 0.5 \end{bmatrix}, \qquad K_y = \begin{bmatrix} -0.5 \\ 0 \\ 0.5 \end{bmatrix}. \tag{4.13}$$

The corresponding finite–difference gradient loss function is then defined as

$$L_{\text{FD}} = \frac{1}{HW} \sum_{i,j} \left(\left| \hat{x}_{x,i,j} - x_{x,i,j} \right| + \left| \hat{x}_{y,i,j} - x_{y,i,j} \right| \right). \tag{4.14}$$

As an alternative to finite differences, spatial gradients can be computed using the Sobel operator (Sobel, Feldman, et al., 1968), with the kernels

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \qquad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \tag{4.15}$$

We define the Sobel–based gradient loss function, L_{SO}, analogously to Equation 4.14, with the finite–difference kernels replaced by the Sobel kernels.

We denote the gradient regularization term generically as $L_{\text{grad}} \in \{L_{\text{FD}}, L_{\text{SO}}\}$. Finally, we combine the reconstruction loss, $L_{\text{rec}} \in \{L_{\text{MSE}}, L_1, L_{\text{SL}_1}\}$, also referred to as the primary loss function, with a gradient–based auxiliary loss, $L_{\text{grad}} \in \{L_{\text{FD}}, L_{\text{SO}}\}$, to obtain the total training objective:

$$L_{\text{total}} = L_{\text{rec}} + \lambda L_{\text{grad}}, \tag{4.16}$$

where λ controls the relative contribution of the gradient–based auxiliary loss. The parameter λ is treated as a hyperparameter rather than a learnable parameter.

The preceding discussion motivates the reconstruction and gradient losses from the perspective of image fidelity and the preservation of spatial structure. These terms also have a direct connection with the physical quantities introduced in Section 4.2.1. For clarity, the relations below are expressed in physical units using the notation introduced in that section. Since the same global linear scaling is applied to the original and reconstructed magnetograms, the corresponding pixel and gradient differences used during training differ only by a constant scale factor.

Consider first the total unsigned LoS magnetic flux,

$$\Phi = \sum_{x,y} |B_{\text{LoS}}(x,y)|\, \Delta A, \tag{4.17}$$

and let $\hat{\Phi}$ denote the same quantity computed from the reconstructed magnetogram $\hat{B}_{\text{LoS}}(x,y)$. The difference between the two quantities satisfies

$$\left|\Phi - \hat{\Phi}\right| = \Delta A \left| \sum_{x,y} \left(|B_{\text{LoS}}(x,y)| - \left|\hat{B}_{\text{LoS}}(x,y)\right| \right) \right|. \tag{4.18}$$

Applying the triangle inequality gives

$$\left|\Phi - \hat{\Phi}\right| \leq \Delta A \sum_{x,y} \left| |B_{\text{LoS}}(x,y)| - \left|\hat{B}_{\text{LoS}}(x,y)\right| \right|, \tag{4.19}$$

and, from the reverse triangle inequality,

$$\left| |B_{\text{LoS}}(x,y)| - \left|\hat{B}_{\text{LoS}}(x,y)\right| \right| \leq \left| B_{\text{LoS}}(x,y) - \hat{B}_{\text{LoS}}(x,y) \right|. \tag{4.20}$$

Therefore,

$$\left|\Phi - \hat{\Phi}\right| \leq \Delta A \sum_{x,y} \left| B_{\text{LoS}}(x,y) - \hat{B}_{\text{LoS}}(x,y) \right|. \tag{4.21}$$

Thus, pixel–wise reconstruction accuracy directly constrains the magnetic–field amplitudes from which the unsigned flux is computed. In the particular case of an L_1 reconstruction objective, the expression above provides an explicit upper bound on the error in Φ. The other pixel–wise losses considered here act on the same local residuals, although they weight their magnitude differently.

A similar, and more direct, connection exists for the mean LoS magnetic–field gradient,

$$\langle|\nabla B_{\mathrm{LoS}}|\rangle = \frac{1}{N}\sum_{x,y}|\nabla B_{\mathrm{LoS}}(x,y)|. \tag{4.22}$$

Let $\langle|\nabla \hat{B}_{\mathrm{LoS}}|\rangle$ denote the corresponding quantity computed from the reconstructed magnetogram. Applying the triangle inequality gives

$$\left|\langle|\nabla B_{\mathrm{LoS}}|\rangle - \langle|\nabla \hat{B}_{\mathrm{LoS}}|\rangle\right| \leq \frac{1}{N}\sum_{x,y}\left||\nabla B_{\mathrm{LoS}}(x,y)| - |\nabla \hat{B}_{\mathrm{LoS}}(x,y)|\right|. \tag{4.23}$$

Using the reverse triangle inequality for the gradient vectors,

$$\left||\nabla B_{\mathrm{LoS}}(x,y)| - |\nabla \hat{B}_{\mathrm{LoS}}(x,y)|\right| \leq \left|\nabla B_{\mathrm{LoS}}(x,y) - \nabla \hat{B}_{\mathrm{LoS}}(x,y)\right|. \tag{4.24}$$

Furthermore,

$$\begin{aligned}&\left|\nabla B_{\mathrm{LoS}}(x,y) - \nabla \hat{B}_{\mathrm{LoS}}(x,y)\right| \\ &\quad\leq \left|\frac{\partial B_{\mathrm{LoS}}}{\partial x} - \frac{\partial \hat{B}_{\mathrm{LoS}}}{\partial x}\right| + \left|\frac{\partial B_{\mathrm{LoS}}}{\partial y} - \frac{\partial \hat{B}_{\mathrm{LoS}}}{\partial y}\right|.\end{aligned} \tag{4.25}$$

The terms on the right–hand side are precisely the type of local derivative discrepancies penalized by L_{FD}. The finite–difference auxiliary loss therefore directly constrains, at corresponding spatial locations, the derivatives from which $\langle|\nabla B_{\mathrm{LoS}}|\rangle$ is calculated.

The relation with R is less direct because this quantity depends not only on the magnetic–field amplitudes but also on the location and geometry of the PIL through the weighting function $p(x,y)$. Pixel–wise fidelity constrains the magnetic amplitudes entering R, while gradient fidelity promotes preservation of the sharp opposite–polarity structures involved in identifying the PIL. However, $p(x,y)$ itself is obtained through a sequence of operations used to identify and spatially weight the PIL, including thresholding and morphological operations. Consequently, the standard computation of R is not directly differentiable with respect to the reconstructed magnetogram. A differentiable approximation could in principle be constructed, but this would introduce an additional methodological choice and would no longer correspond exactly to the diagnostic used here for evaluation.

Although quantities such as Φ and $\langle|\nabla B_{\mathrm{LoS}}|\rangle$ could themselves be incorporated into the objective as differentiable loss terms, doing so would constrain only their final global or spatially averaged values. It is important to distinguish this from the pixel–wise and gradient losses used here. In either formulation, the final quantity minimized during training is a single scalar value; however, the scalar is constructed differently.

For example, a loss defined directly from the total unsigned flux as

$$L_{\Phi} = \left|\Phi - \hat{\Phi}\right| \tag{4.26}$$

would first reduce each magnetogram to a single value, Φ or $\hat{\Phi}$, and only then compare the two. Consequently, two magnetograms could yield the same total unsigned flux, and therefore $L_\Phi = 0$, even if the magnetic flux were distributed differently across the AR.

Similarly, a loss defined directly from the mean LoS magnetic-field gradient would compare only the two spatially averaged gradient values. It could therefore be small even if strong gradients were reconstructed at incorrect spatial locations.

By contrast, the reconstruction and gradient losses adopted here compare the original and reconstructed magnetograms at corresponding spatial locations before the individual discrepancies are aggregated into the final scalar loss. For L_{rec}, the residual is evaluated independently at each pixel, while for L_{FD}, the horizontal and vertical derivatives are compared locally at each corresponding position. Therefore, an error in the spatial location of magnetic flux or strong gradients contributes to the loss even when the corresponding global physical quantity is preserved. The final optimization objective is still scalar, but it retains spatial information through the local comparisons from which that scalar is constructed.

For this reason, the physical quantities are retained as independent evaluation diagnostics rather than explicitly included in the optimization objective. This allows us to assess whether improving the reconstruction of the magnetic field and its local spatial derivatives also leads to improved physical consistency without directly enforcing the evaluated quantities during training.

Models are trained for 100 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 128. Model checkpoints are saved at each epoch and monitored according to the validation loss, retaining the configurations with the lowest validation loss values once the loss falls below the initial threshold of 0.01. To improve training stability, the auxiliary loss is activated only after an initial warm–up period (the first five epochs), during which training is driven solely by L_{rec}. The weighting coefficient λ is selected via a grid search over logarithmically spaced values, $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, with the final configuration chosen based on validation performance.

4.3.3 Image Reconstruction Evaluation Tools

To assess model performance, we compute a set of complementary image–based metrics that quantify (i) pixel–wise residual errors, (ii) structural similarity, and (iii) distributional agreement between the original magnetogram $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$. During test–set image reconstruction, the preprocessing pipeline described in Section 4.2.4 is applied with center cropping instead of random cropping. All metrics are computed for each original–reconstruction pair and averaged over the test set.

To account for pixel–wise reconstruction error, we report the *mean squared error (MSE)*,

$$MSE(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(x_{i,j} - \hat{x}_{i,j}\right)^2, \tag{4.27}$$

and the *mean absolute error (MAE)*,

$$MAE(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left|x_{i,j} - \hat{x}_{i,j}\right|. \tag{4.28}$$

These two metrics provide a first quantitative assessment of reconstruction performance. However, pixel–wise metrics treat each residual independently and therefore do not explicitly encode spatial organization or correlations (Wang and Bovik, 2009).

To assess structural fidelity beyond pixel–wise errors, we compute the *structural similarity index (SSIM*; Wang et al., 2004) between $\mathbf{x}$ and $\hat{\mathbf{x}}$. The *SSIM* is computed locally by evaluating, at each spatial location j, the similarity between local patches $\mathbf{x}_j$ and $\hat{\mathbf{x}}_j$ centered at j,

$$SSIM(\mathbf{x}_j, \hat{\mathbf{x}}_j) = \frac{(2\mu_{x,j}\mu_{\hat{x},j} + C_1)(2\sigma_{x\hat{x},j} + C_2)}{(\mu_{x,j}^2 + \mu_{\hat{x},j}^2 + C_1)(\sigma_{x,j}^2 + \sigma_{\hat{x},j}^2 + C_2)}, \tag{4.29}$$

where $\mu_{(\cdot),j}$ denotes the local mean, $\sigma_{(\cdot),j}^2$ the local variance, and $\sigma_{x\hat{x},j}$ the local covariance. The constants C_1 and C_2 stabilize the expression in low–contrast regions and follow the standard choice $C_1 = (0.01L)^2$ and $C_2 = (0.03L)^2$, with the dynamic range taken as $L = \max(\mathbf{x}) - \min(\mathbf{x})$ for each magnetogram.

SSIM compares local statistics over spatial neighborhoods, assessing whether the reconstruction preserves local intensity (mean), contrast (variance), and structure (covariance), rather than merely matching individual pixel values. Following standard implementations, we compute an *SSIM* map using a 7×7 square window and report the *mean structural similarity index (MSSIM)*, defined as the mean of the local *SSIM* values over all valid window locations,

$$MSSIM(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{M} \sum_{j=1}^{M} SSIM(\mathbf{x}_j, \hat{\mathbf{x}}_j). \tag{4.30}$$

Autoencoder reconstructions may preserve local contrast and spatial correlation patterns, so that *MSSIM* remains high and the image appears qualitatively similar, while still altering the global distribution of pixel intensities. To assess distributional agreement, we compare the normalized pixel–intensity histograms P and Q of $\mathbf{x}$ and $\hat{\mathbf{x}}$ using the *Kullback–Leibler (KL) divergence* (Kullback and Leibler, 1951),

$$D_{KL}(P \| Q) = \sum_k P(k) \log\left(\frac{P(k)}{Q(k)}\right). \tag{4.31}$$

The *KL divergence* quantifies the information loss incurred when the reconstructed pixel–intensity distribution Q is used to approximate the original pixel–intensity distribution P. It is therefore sensitive to systematic mismatches such as shifts in the overall intensity distribution or changes in the relative frequency of weak and strong field values. Histograms are computed from the flattened pixel values using a common set of bin edges, here 500 bins, and are normalized to form discrete probability distributions.

To assess whether reconstructions preserve structure across spatial scales, we compare the Fourier power spectra of $\mathbf{x}$ and $\hat{\mathbf{x}}$. We subtract the mean from each image and apply a separable Hann window prior to computing the Fourier transform. The two–dimensional power spectrum of an image $\mathbf{x}$ is defined as

$$P_{\mathbf{x}}(k_x, k_y) = \left|\mathcal{F}\{\mathbf{x}\}(k_x, k_y)\right|^2, \tag{4.32}$$

where $\mathcal{F}\{\mathbf{x}\}(k_x, k_y)$ denotes the two–dimensional Fourier transform evaluated at the discrete Fourier modes (k_x, k_y). The same definition applies to $\hat{\mathbf{x}}$. To obtain a rotation–invariant representation and facilitate comparison across magnetograms, we compute the azimuthally averaged radial power spectrum $P(k)$ by averaging $P(k_x, k_y)$ over concentric annuli in Fourier space.

The radial power spectrum summarizes the relative contribution of different spatial scales and enables a direct comparison of scale–dependent fidelity between the original and reconstructed images, making it possible to detect both smoothing (a deficit of high–frequency power) and mismatched fine–scale texture (a redistribution of power at high frequencies) even when pixel–wise errors remain small. The resulting spectra are visualized on a log–log scale up to the Nyquist frequency, taken here as $\min(H, W)/2$.

4.3.4 Time–Series Evaluation Tools

Beyond pixel–level fidelity, we assess whether reconstructed LoS magnetograms preserve physically meaningful parameters over time. We therefore compute three parameters derived from each LoS magnetogram in the test set: the total unsigned LoS magnetic flux Φ (Equation 4.1), the gradient–weighted neutral line length R (Equation 4.2), and the mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$ (Equation 4.3).

For each AR in the test set, these physical parameters are computed at every time step from both the original magnetograms $\mathbf{x}$ and the reconstructed magnetograms $\hat{\mathbf{x}}$. For each parameter, this produces paired time series $\{p_t\}_{t=1}^{T}$ and $\{\hat{p}_t\}_{t=1}^{T}$, computed from $\mathbf{x}$ and $\hat{\mathbf{x}}$, respectively, which track its temporal evolution over the lifetime of the AR. We evaluate agreement between the original–derived and reconstruction–derived parameter time series using complementary metrics that capture absolute deviation, trend similarity, shape similarity, and temporal alignment.

To quantify pointwise discrepancies, we report the *mean absolute error (MAE)*

$$MAE(p, \hat{p}) = \frac{1}{T} \sum_{t=1}^{T} |p_t - \hat{p}_t| \,. \tag{4.33}$$

This metric provides an interpretable error scale in the native physical units of the parameter. Because the physical parameters considered in this work have different units and dynamic ranges, raw *MAE* values are not directly comparable across parameters. To enable cross–parameter comparison, we additionally report the *standardized mean absolute error (sMAE)* obtained by normalizing with the standard deviation of the original–derived parameter series. Let

$$\bar{p} = \frac{1}{T} \sum_{t=1}^{T} p_t, \qquad \sigma_p = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (p_t - \bar{p})^2}, \tag{4.34}$$

we then define the *sMAE* as

$$sMAE(p, \hat{p}) = \frac{1}{T} \sum_{t=1}^{T} \frac{|p_t - \hat{p}_t|}{\sigma_p}. \tag{4.35}$$

The *sMAE* is dimensionless and expresses reconstruction error relative to the intrinsic variability of the parameter time series, allowing meaningful comparison of reconstruction performance across different parameters.

MAE and *sMAE* do not indicate whether the reconstruction–derived parameter time series reproduces the temporal co–variation of the original–derived parameter time series. We therefore evaluate linear similarity using the *Pearson correlation coefficient* (Pearson, 1895),

$$\rho = \frac{\mathrm{cov}(p, \hat{p})}{\sigma_p \, \sigma_{\hat{p}}}, \tag{4.36}$$

where

$$\mathrm{cov}(p, \hat{p}) = \frac{1}{T} \sum_{t=1}^{T} (p_t - \bar{p})(\hat{p}_t - \bar{\hat{p}}), \tag{4.37}$$

and

$$\sigma_p = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (p_t - \bar{p})^2}, \qquad \sigma_{\hat{p}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\hat{p}_t - \bar{\hat{p}})^2}. \tag{4.38}$$

The *Pearson correlation coefficient* is scale–invariant and indicates whether the reconstruction–derived parameter time series preserves the global trend and relative variability of the original–derived parameter time series, even in the presence of small amplitude biases. However, *Pearson correlation coefficient* quantifies linear co–variation at zero lag and does not distinguish discrepancies caused by differences in temporal shape (e.g., sequence of rises, peaks, and decays) from those caused by temporal misalignment, in which

similar features occur at different times (e.g., due to systematic advances or delays).

To characterize shape similarity beyond linear co–variation, we compute the *dynamic time warping (DTW) distance* (Sakoe and Chiba, 1978). DTW compares two time series by allowing local stretching or compression of the time axis. Let $D(i,j)$ denote the minimum cumulative cost required to align the first i points of the original–derived parameter series $\{p_t\}_{t=1}^{T}$ with the first j points of the reconstruction–derived parameter series $\{\hat{p}_t\}_{t=1}^{T}$. Using the local absolute difference
the standard DTW recursion is

$$D(i,j) = c(i,j) + \min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\}. \tag{4.39}$$

The *DTW distance* is then defined as the final cumulative alignment cost,

$$DTW(p,\hat{p}) = D(T,T). \tag{4.40}$$

DTW identifies an optimal monotone warping path that aligns the two sequences in time, thereby providing a similarity measure that is robust to local phase shifts and variable temporal rates.

Because the *DTW distance* is a cumulative alignment cost, its value depends on both the scale of the physical parameter and the length of the time series. To enable comparison across parameters and ARs, we also report a *standardized DTW distance*, *sDTW distance*. Similar to *sMAE*, *sDTW distance* is normalized by the standard deviation of the original–derived parameter series; in addition, it is divided by the sequence length T,

$$sDTW(p,\hat{p}) = \frac{DTW(p,\hat{p})}{T\,\sigma_p}. \tag{4.41}$$

This normalized quantity is dimensionless and expresses the cumulative shape mismatch relative to both the intrinsic variability of the original–derived parameter series and the number of time steps.

DTW provides a flexible measure of shape similarity under non–uniform time warping, but it does not directly quantify a systematic lead or lag between two time series. To explicitly measure such temporal offsets, we evaluate temporal alignment using the cross–correlation function (CCF; Box et al., 2015), which measures linear similarity as a function of an imposed global time shift k. For an integer lag k, we define

$$CCF(k) = \frac{\sum_{t\in\mathcal{T}_k}(p_t-\bar{p})\,(\hat{p}_{t-k}-\bar{\hat{p}})}{\sqrt{\sum_{t\in\mathcal{T}_k}(p_t-\bar{p})^2}\,\sqrt{\sum_{t\in\mathcal{T}_k}(\hat{p}_{t-k}-\bar{\hat{p}})^2}}, \tag{4.42}$$

where $\mathcal{T}_k$ denotes the set of indices for which both p_t and $\hat{p}_{t-k}$ are defined (i.e., the overlapping time interval at lag k). We report the *CCF peak lag*,

$$k^\star = \arg\max_k \mathrm{CCF}(k), \tag{4.43}$$

which estimates a systematic advance or delay of the reconstructed series relative to the reference. Unlike DTW, CCF does not permit local time warping: the shift k is uniform across the entire series. A zero *CCF peak lag* indicates correct temporal alignment, whereas a non–zero *CCF peak lag* reveals a systematic time shift in the evolution of the physical parameter recovered from the reconstructed magnetograms.

4.4 Results

This section presents the empirical evaluation of the proposed CAE for the reconstruction of AR LoS magnetograms. We compare the performance of the same CAE architecture when trained with different training objectives. Specifically, we evaluate reconstruction–only baselines, as well as composite training objectives that combine the primary reconstruction losses with auxiliary losses designed to promote the preservation of spatial structure in the pixel space.

The best–performing configurations are selected based on validation–set performance, considering both the reconstruction–only baselines and the composite training objectives associated with each primary reconstruction loss. The analysis is guided by two main questions: (1) do the selected training objectives improve the overall test–set image reconstruction quality? and (2) do improvements in image–level reconstruction translate into physically consistent temporal evolution of derived magnetic–field parameters?

To answer these questions, we conduct a controlled comparison of the different configurations using the same dataset split, training setup, and evaluation protocol. In the following sections, we present quantitative and qualitative analyses that assess reconstruction accuracy, structural and spectral fidelity, and the preservation of derived physical quantities across time.

4.4.1 Model Selection

The CAE architecture and the candidate training objectives have been defined in Sections 4.3.1 and 4.3.2, respectively. The next step is to select the best model configuration using only the validation set. Throughout this stage, the CAE architecture is kept fixed, so that model selection concerns only the choice of the best training objective. In particular, the model is trained either with a reconstruction–only objective, used as a baseline, or with a composite training objective of the form

$$L_{\text{total}} = L_{\text{rec}} + \lambda L_{\text{grad}},$$

where $L_{\text{rec}} \in \{L_{\text{MSE}}, L_1, L_{\text{SL}_1}\}$ denotes the primary reconstruction loss, $L_{\text{grad}} \in \{L_{\text{FD}}, L_{\text{SO}}\}$ denotes the gradient–based auxiliary loss, and λ controls the relative contribution of the auxiliary term. The validation results for all candidate configurations are reported in Table 4.4. The baseline column corresponds to models trained without an auxiliary loss, while the remaining columns correspond to different values of λ.

Primary Loss	Auxiliary Loss	Baseline	λ			
			10^{-2}	10^{-3}	10^{-4}	10^{-5}
L_{MSE}	None	9.230×10^{-6}	–	–	–	–
	L_{FD}	–	3.820×10^{-6}	4.102×10^{-6}	5.688×10^{-6}	5.979×10^{-6}
	L_{SO}	–	3.920×10^{-5}	1.499×10^{-5}	1.123×10^{-5}	1.058×10^{-5}
L_1	None	1.480×10^{-5}	–	–	–	–
	L_{FD}	–	1.543×10^{-5}	6.946×10^{-6}	6.644×10^{-6}	5.151×10^{-6}
	L_{SO}	–	5.836×10^{-6}	1.416×10^{-5}	5.722×10^{-6}	7.876×10^{-6}
L_{SL_1}	None	1.126×10^{-5}	–	–	–	–
	L_{FD}	–	2.673×10^{-6}	7.321×10^{-6}	8.698×10^{-6}	9.817×10^{-6}
	L_{SO}	–	5.617×10^{-5}	7.353×10^{-6}	7.855×10^{-6}	4.232×10^{-6}

TABLE 4.4: Validation *MSE* for the candidate autoencoder configurations. Rows indicate the primary reconstruction loss and the auxiliary loss, while columns report the auxiliary–loss weight λ. The baseline column reports configurations trained without an auxiliary loss.

All configurations are compared using a common validation metric. Specifically, model selection is based on the validation *MSE* (Equation 4.27), computed between the input LoS magnetograms and their reconstructions. Using *MSE* avoids directly comparing loss values across heterogeneous training objectives. The test set is not used during model selection and is reserved for the final evaluation of the selected configurations.

Based on the validation results, three configurations are retained for the final test–set evaluation: the best reconstruction–only baseline, the best model with finite–difference gradient regularization, and the best model with Sobel–based gradient regularization. The selected configurations are reported in Table 4.5. The lowest validation MSE is obtained by the configuration trained with L_{SL_1} as the primary loss and L_{FD} as the auxiliary loss, with $\lambda = 10^{-2}$. This configuration achieves a validation MSE of 2.673×10^{-6}.

Primary Loss	Auxiliary Loss	λ	Validation MSE
L_{MSE}	None	–	9.230×10^{-6}
L_{SL_1}	L_{SO}	10^{-5}	4.232×10^{-6}
L_{SL_1}	L_{FD}	10^{-2}	2.673×10^{-6}

TABLE 4.5: Selected autoencoder configurations retained for the final test–set evaluation. The table reports the best reconstruction–only baseline, the best finite–difference gradient–regularized model, and the best Sobel–regularized model according to validation MSE.

4.4.2 Image Reconstruction Evaluation

Following the validation–based model selection performed in the previous section, the selected models were evaluated on the independent test set. These models share the same underlying CAE architecture and differ only in the training objective. Hereafter, the models are referred to according to the objective used during training: the reconstruction–only baseline model, the finite–difference gradient–regularized model, and the Sobel–regularized model. This final evaluation provides a common quantitative comparison of how different training objectives affect image reconstruction quality. Four complementary metrics are used. The *MSE* and *MAE* measure pixel–wise reconstruction errors. The *MSSIM* measures structural agreement between the original and reconstructed LoS magnetograms. The *KL divergence* measures the discrepancy between their pixel–value distributions.

The test–set results presented in Table 4.6 are consistent with the validation–based selection. Among the evaluated models, the finite–difference gradient–regularized model, $L_{\mathrm{SL}_1} + 10^{-2} L_{\mathrm{FD}}$, achieved the best reconstruction performance in terms of pixel–wise and structural metrics, with the lowest *MSE* (3.000×10^{-6}), the lowest *MAE* (9.160×10^{-4}), and the highest *MSSIM* ($9.967 \times$

Primary loss	Auxiliary loss	λ	MSE ↓	MAE ↓	MSSIM ↑	KL ↓
L_{MSE}	None	–	1.100×10^{-5}	1.524×10^{-3}	9.897×10^{-1}	3.001×10^{-2}
L_{SL_1}	L_{SO}	10^{-5}	5.000×10^{-6}	1.158×10^{-3}	9.952×10^{-1}	9.126×10^{-2}
L_{SL_1}	L_{FD}	10^{-2}	3.000×10^{-6}	9.160×10^{-4}	9.967×10^{-1}	1.119×10^{-1}

TABLE 4.6: Image–level reconstruction metrics computed on the test set for the selected autoencoder configurations. The table reports pixel–wise reconstruction errors, structural similarity, and distributional agreement between the original and reconstructed LoS magnetograms.

10^{-1}). The Sobel–regularized model, $L_{\text{SL}_1} + 10^{-5} L_{\text{SO}}$, also improved over the reconstruction–only baseline in these metrics, although it did not match the performance of the finite–difference configuration. The only exception was the pixel–value distribution metric, for which the reconstruction–only baseline L_{MSE} gave the lowest *KL divergence* (3.001×10^{-2}).

The lower *KL divergence* obtained with the reconstruction–only objective may be related to the quadratic dependence of L_{MSE} on the reconstruction residuals. Since large residuals are more strongly penalized, training with L_{MSE} may favor preservation of the global amplitude distribution of the magnetogram, including less frequent high–intensity values. This can lead to a reconstructed pixel–value distribution that is closer to the original distribution, even when the reconstruction is less accurate in terms of local structure. By contrast, the gradient–based objectives improve spatial sharpness and structural agreement, but may redistribute errors in a way that slightly alters the global histogram, producing larger *KL* divergence values.

To complement the aggregate test–set analysis, we next examine two representative examples in more detail. This qualitative analysis is intended to illustrate how the selected models behave when reconstructing magnetograms from different magnetic contexts. We consider two contrasting examples: HARP 377, observed on 2011.02.15 at 02:00 during an X2.2 flare, and HARP 8906, observed on 2022.12.21 at 12:48 and not associated with a flare. These examples provide a visual comparison between a complex flare–productive AR and a quieter non–flaring one.

Figure 4.2 compares the radially averaged Fourier power spectra of the original magnetograms and the reconstructions obtained with the selected models for these two examples. In both cases, the spectra of all reconstructions overlap closely with the original spectrum at low and intermediate values of k, indicating that the large–scale and intermediate–scale magnetic structures are well preserved independently of the training objective.

Differences become more evident at larger values of k, as shown in the zoomed panels. Since the radial Fourier mode k is inversely related to spatial scale, larger values of k correspond to progressively finer structures. In this

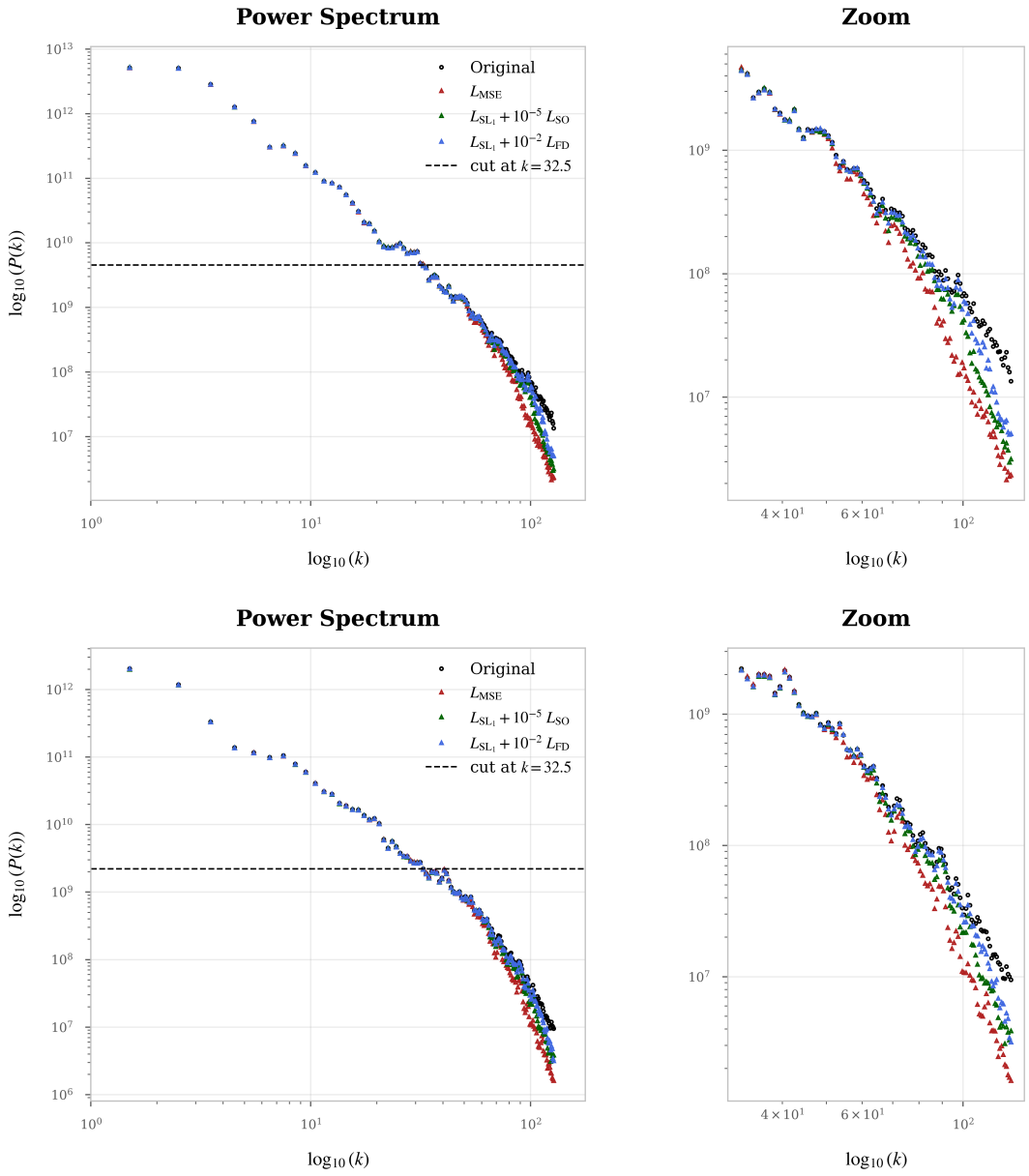


FIGURE 4.2: Comparison of the radially averaged Fourier power spectra for the original LoS magnetograms and the reconstructions obtained with the selected models. The two examples correspond to HARP 377 (upper panels), observed on 2011.02.15 at 02:00 during an X2.2 flare, and HARP 8906 (lower panels), observed on 2022.12.21 at 12:48 and not associated with a flare. The left panels show the full spectra, while the right panels show a zoom of the high–k range. The dashed horizontal line marks the power value at $k = 32.5$, which is used as a reference for the high–k range shown in the zoomed panels.

regime, the reconstruction–only baseline shows the strongest attenuation of power relative to the original spectrum, indicating a greater loss of fine–scale structure. The gradient–regularized models retain more power at high k and remain closer to the original spectrum, suggesting that the auxiliary gradient losses help preserve small–scale magnetic variations.

Figures 4.3 and 4.4 show representative reconstruction examples obtained with the finite–difference gradient–regularized model, which achieved the best test–set performance in terms of pixel–wise and structural metrics. The two examples considered are the same as above: the complex flare–productive AR HARP 377 and the quieter non–flaring AR HARP 8906. In both cases, the large–scale magnetic morphology is accurately recovered, including the spatial arrangement of the dominant positive and negative polarity concentrations. Small–scale positive– and negative–polarity flux concentrations and sharp transitions are also reproduced, although their boundaries appear slightly smoothed in the reconstructed magnetograms.

The polarity–separated residual maps show the absolute reconstruction error independently in positive– and negative–field regions. In both examples, the residuals are spatially localized and concentrate mainly near compact flux concentrations, strong–field gradients, and PILs. These regions are characterized by sharp magnetic transitions; as a result, even small spatial offsets between the original and reconstructed magnetograms can produce relatively large local residuals. Nevertheless, the residual amplitudes remain much smaller than the peak magnetic–field values shown in the corresponding magnetograms, indicating that the errors are local perturbations rather than large–scale morphological failures. This qualitative analysis of the reconstructions and residual maps is consistent with the low *MSE* and *MAE*, and the high *MSSIM*, reported in Table 4.6.

The pixel–value distributions provide a qualitative check of the agreement between the original and reconstructed magnetic–field amplitudes. For both examples, the reconstructed distributions closely follow the original ones over most of the dynamic range, including the central background component and the extended tails associated with stronger magnetic–field concentrations. Small deviations are visible in the distribution tails, consistent with the larger *KL divergence* obtained for the finite–difference gradient–regularized model. This indicates that the finite–difference objective improves structural preservation while still allowing slight redistributions of magnetic–field amplitudes. In particular, the strongest magnetic–field amplitudes may be slightly underrepresented or redistributed toward intermediate values.

Figures 4.5 and 4.6 provide a local visual comparison of the reconstructions produced by the three selected models in regions containing compact magnetic structure and sharp transitions. In both examples, the reconstruction–only baseline reproduces the broad magnetic pattern but shows larger absolute deviations. The Sobel–regularized and finite–difference gradient–regularized models produce lower residual amplitudes in the selected regions and more closely follow the original local morphology, consistent with the conclusions drawn from the aggregate test–set metrics and the power–spectrum analysis.

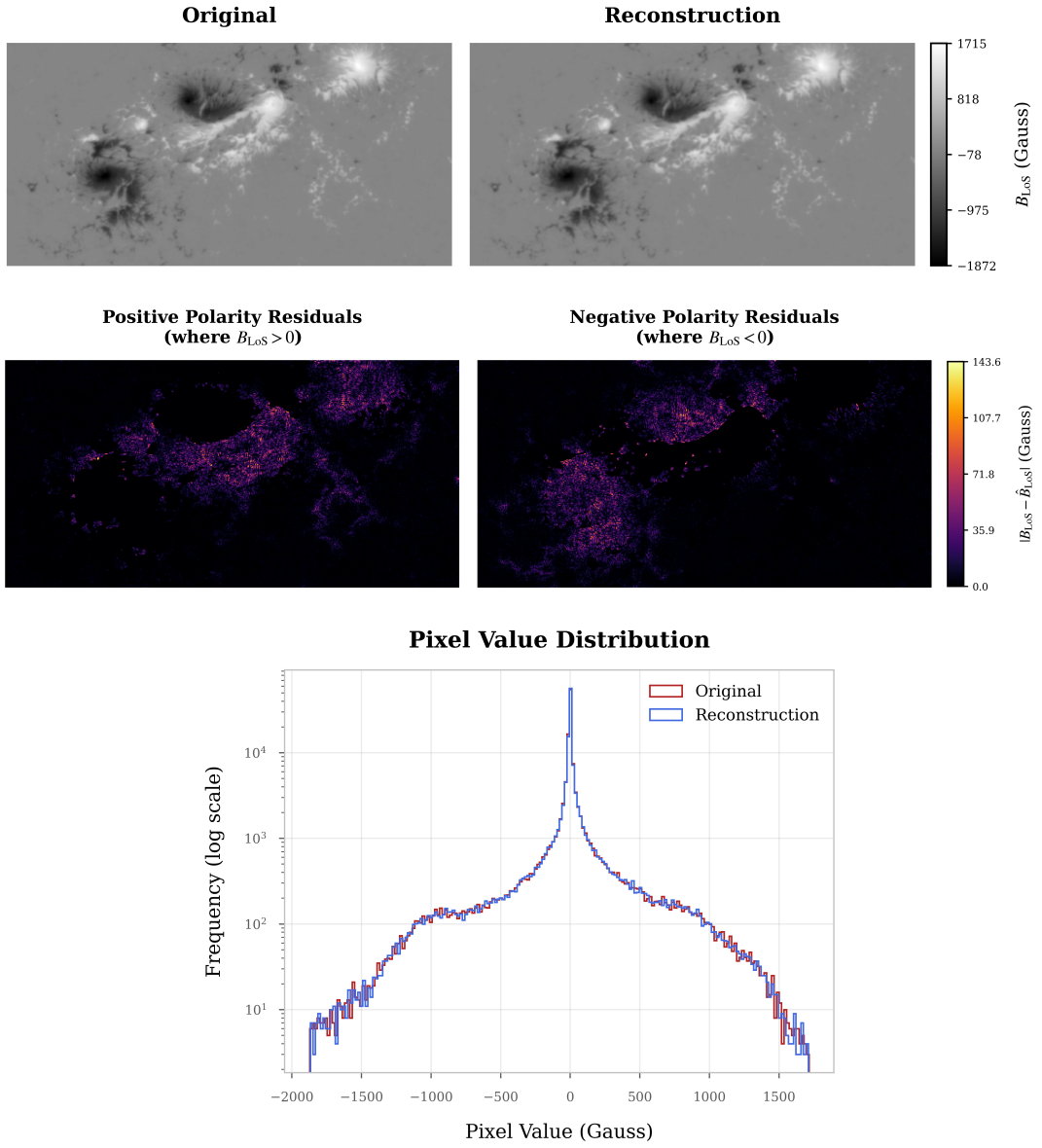


FIGURE 4.3: Representative reconstruction example for HARP 377, observed on 2011.02.15 at 02:00 during an X2.2 flare. The reconstruction is obtained with the finite–difference gradient–regularized model. The panels show the original and reconstructed LoS magnetograms, polarity–separated absolute residual maps, and the pixel–value distributions.

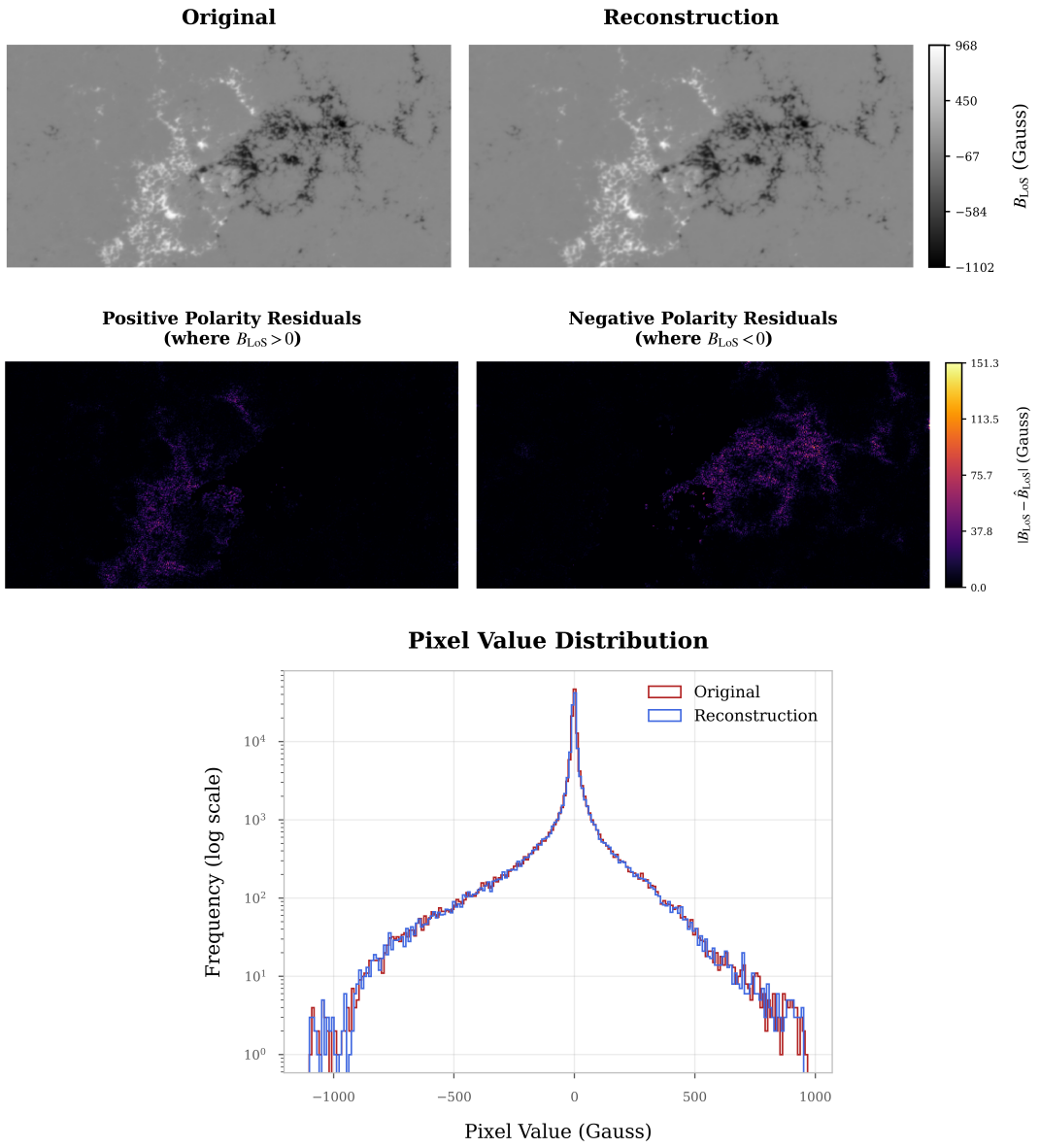


FIGURE 4.4: Representative reconstruction example for HARP 8906, observed on 2022.12.21 at 12:48 and not associated with a flare. The reconstruction is obtained with the finite–difference gradient–regularized model. The panels show the original and reconstructed LoS magnetograms, polarity–separated absolute residual maps, and the pixel–value distributions.

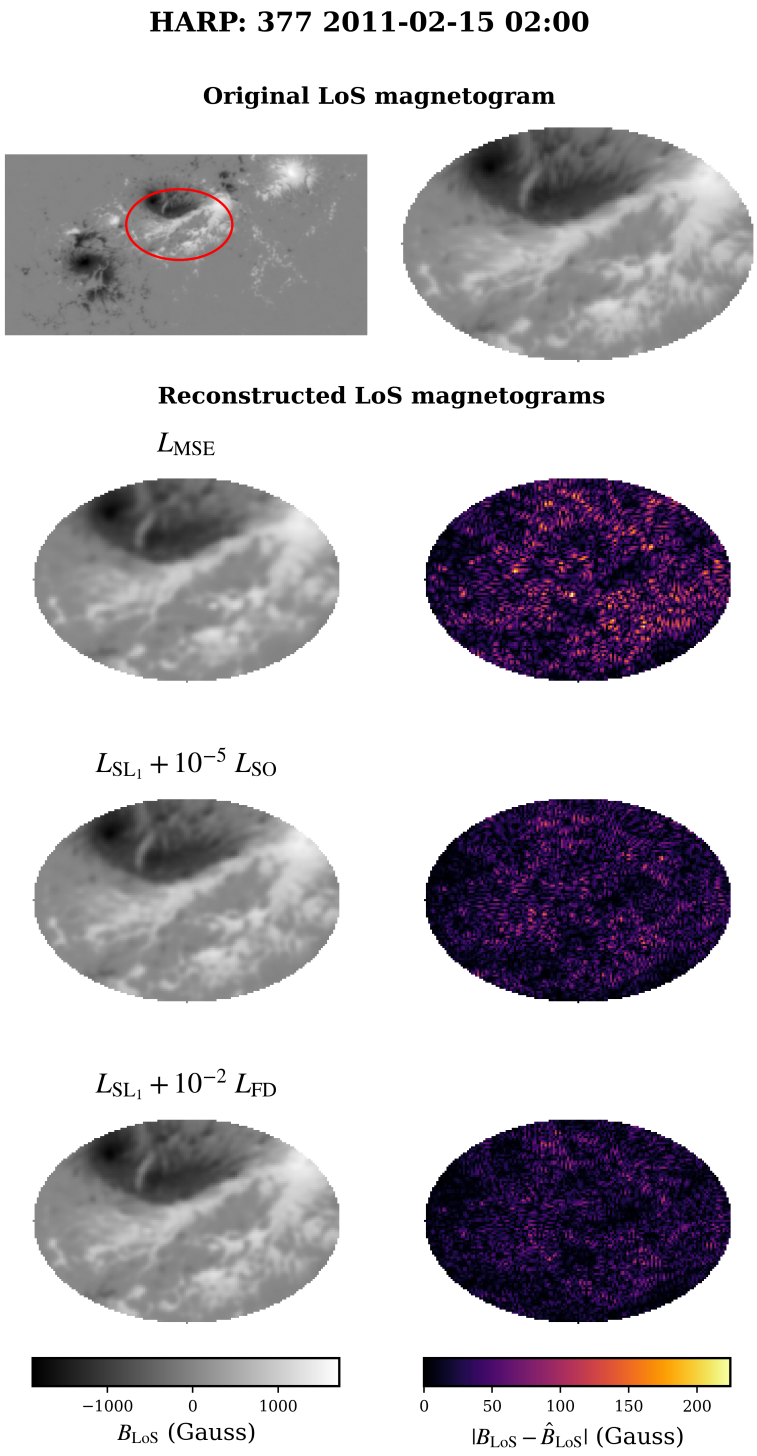


FIGURE 4.5: Local reconstruction comparison for HARP 377, observed on 2011.02.15 at 02:00 during an X2.2 flare. The red ellipse in the full LoS magnetogram identifies the selected region. The figure compares the selected original region with the reconstructions obtained using the reconstruction–only baseline model, the Sobel–regularized model, and the finite–difference gradient–regularized model, together with their absolute deviation maps.

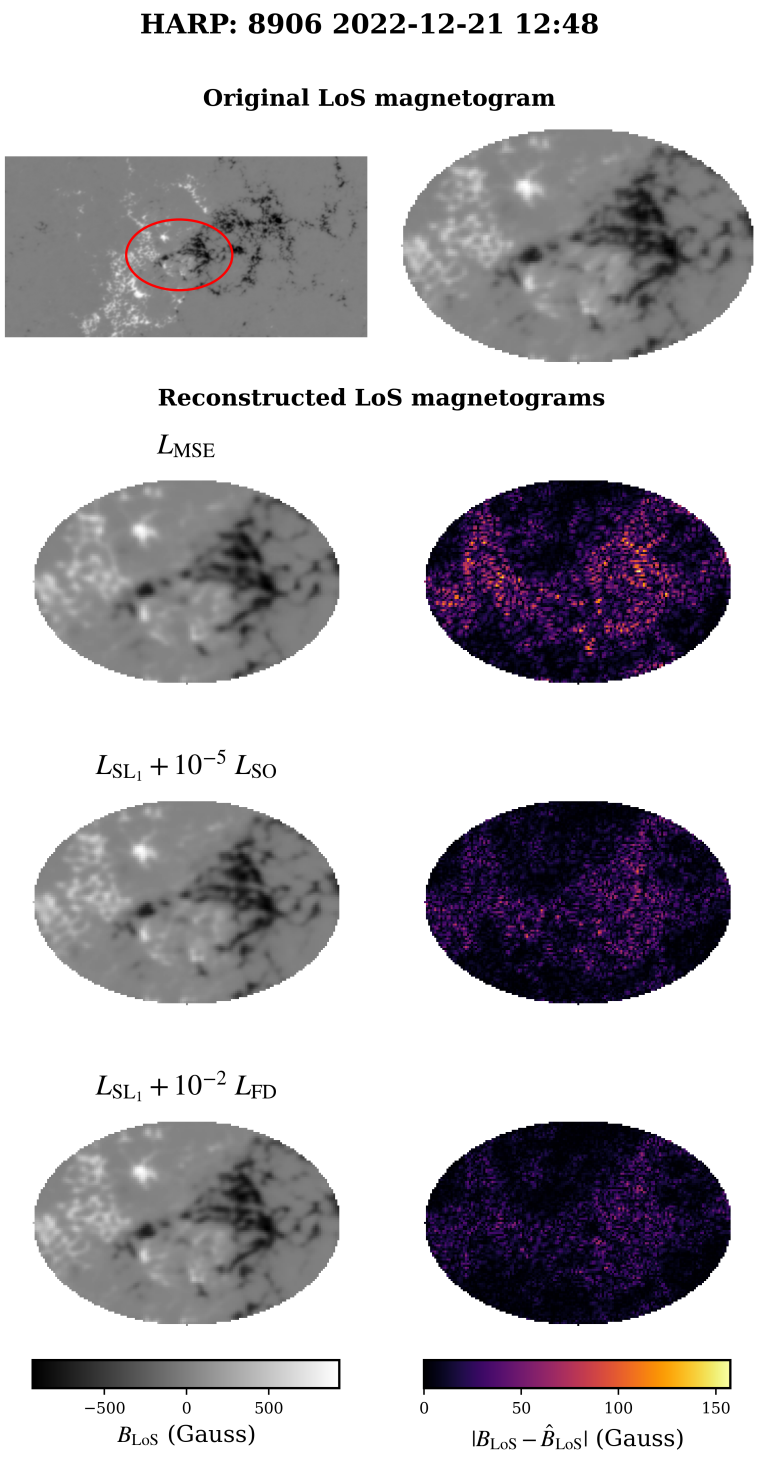


FIGURE 4.6: Local reconstruction comparison for HARP 8906, observed on 2022.12.21 at 12:48 and not associated with a flare. The red ellipse in the full LoS magnetogram identifies the selected region. The figure compares the selected original region with the reconstructions obtained using the reconstruction–only baseline model, the Sobel–regularized model, and the finite–difference gradient–regularized model, together with their absolute deviation maps.

Overall, the finite–difference gradient–regularized model provides the strongest reconstruction performance among the selected models, particularly in terms of pixel–wise accuracy, structural similarity, and preservation of spectral power at high values of k. The reconstruction–only baseline remains more favorable for the global pixel–value distribution, while the Sobel–regularized model generally shows intermediate behavior.

4.4.3 Temporal Evolution of Derived Physical Parameters

In the previous subsection, model performance is evaluated using image–level reconstruction metrics. Here, we assess whether the reconstructed magnetograms preserve physically meaningful quantities, as well as their temporal evolution. This step is important because small image–level reconstruction errors may propagate differently into integrated or gradient–sensitive physical parameters. The analysis focuses on the total unsigned LoS magnetic flux Φ, the gradient–weighted neutral line length R, and the mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$. For each AR in the test set, these parameters are computed from both the original and reconstructed magnetograms. The comparison is based on the temporal metrics introduced in Section 4.3.4, and the aggregate results are reported in Table 4.7.

The temporal evolution of the total unsigned LoS magnetic flux Φ is preserved with high accuracy across the selected models. All three models yield *Pearson correlation coefficient* values close to unity and zero *CCF peak lag*, indicating that the reconstructed magnetograms reproduce the temporal trend of the unsigned flux without introducing a systematic time shift. Differences between the three models become more evident when examining the standardized error metrics. The best error–based performance is obtained with the finite–difference gradient–regularized model, which gives the lowest *sMAE* (2.262×10^{-2}) and the lowest *sDTW distance* (5.760×10^{-4}). The Sobel–regularized model also improves over the reconstruction–only baseline, although it does not match the finite–difference model. This result indicates that the selected gradient–regularized models preserve the unsigned magnetic flux more accurately and more consistently over time than the reconstruction–only baseline.

For the gradient–weighted neutral line length R, the best performance is again obtained with the finite–difference gradient–regularized model. This configuration gives the lowest *sMAE* (4.404×10^{-2}), the lowest *sDTW distance* (4.268×10^{-3}), and the highest correlation ($\rho = 9.754 \times 10^{-1}$). All three models exhibit comparable *CCF peak lag* values, with a small negative lag of approximately two time steps. This indicates a slight systematic temporal shift between the R time series computed from the original and reconstructed magnetograms. Compared to those of Φ, the *Pearson correlation coefficient* values depart more from unity; however, they remain high across all selected models, indicating that the overall temporal trend of R is preserved. We also note that *sMAE* and *sDTW distance* are higher than for Φ across all three models, indicating that the magnetic features on which the computation of R depends are generally harder to reconstruct. The R parameter is computed

from magnetic flux near PILs, which are localized regions characterized by strong fields, sharp gradients, and high magnetic complexity. These are precisely the structures that are more difficult to reconstruct accurately; therefore, slight smoothing or spatial offsets can more strongly affect the recovery of R and its temporal evolution. In this case as well, the Sobel–regularized model provides an intermediate improvement over the reconstruction–only baseline.

The mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$ shows the clearest benefit from gradient regularization. The finite–difference gradient–regularized model achieves the lowest *sMAE* (2.757×10^{-1}), the lowest *sDTW distance* (4.162×10^{-3}), the highest *Pearson correlation coefficient* ($\rho = 9.992 \times 10^{-1}$), and zero *CCF peak lag*. Compared with the reconstruction–only baseline, this corresponds to a reduction of the standardized amplitude error from 6.542×10^{-1} to 2.757×10^{-1}, and a reduction of the standardized *DTW distance* from 1.064×10^{-2} to 4.162×10^{-3}. This behavior is expected because this parameter is directly determined by local spatial derivatives of the magnetic field; therefore, introducing a regularization term that explicitly constrains local gradients has a direct effect on the recovery of this parameter. The Sobel–regularized model also improves over the reconstruction–only baseline, with $sMAE = 3.597 \times 10^{-1}$ and $sDTW = 5.268 \times 10^{-3}$, but remains below the finite–difference model. These results indicate that both the amplitude and the temporal evolution of magnetic–field gradients are more accurately preserved when local spatial derivatives are explicitly constrained during training.

Figures 4.7 and 4.8 show representative examples of the temporal evolution of the derived physical parameters. The examples correspond to HARP 377 and HARP 8589, which provide two cases with different magnetic evolution and flare activity. In both cases, the parameter time series computed from the reconstructed magnetograms reproduce the main temporal behavior of Φ, R, and $\langle|\nabla B_{\mathrm{LoS}}|\rangle$, confirming that the selected models preserve not only the image morphology, but also the evolution of physically meaningful quantities.

The agreement is strongest for the total unsigned flux Φ, whose reconstructed curves closely follow the values computed from the original magnetograms throughout the observing interval. This is consistent with the integrated nature of Φ, which makes it less sensitive to localized reconstruction errors. Larger differences appear for R, and especially for $\langle|\nabla B_{\mathrm{LoS}}|\rangle$, where the curves show clearer separations between the selected models. This behavior is expected because these quantities depend more directly on localized PIL structures and local magnetic–field gradients, which are more difficult to reconstruct accurately. Consistently with the aggregate results in Table 4.7, the finite–difference gradient–regularized model shows the best overall behavior across the three physical parameters, while the Sobel–regularized model shows intermediate behavior between the reconstruction–only baseline and the finite–difference model.

Physical parameter	Primary loss	Auxiliary loss	λ	MAE ↓	sMAE ↓	DTW ↓	sDTW ↓	ρ ↑	CCF Peak Lag
Φ	L_{MSE}	None	–	6.957×10^{19}	4.382×10^{-2}	8.458×10^{20}	7.740×10^{-4}	1.000	0.000
	L_{SL_1}	L_{SO}	10^{-5}	4.599×10^{19}	3.392×10^{-2}	6.238×10^{20}	7.560×10^{-4}	9.999×10^{-1}	0.000
	L_{SL_1}	L_{FD}	10^{-2}	3.511×10^{19}	2.262×10^{-2}	5.249×10^{20}	5.760×10^{-4}	1.000	0.000
R	L_{MSE}	None	–	6.380×10^{-2}	6.690×10^{-2}	3.946	4.990×10^{-3}	9.610×10^{-1}	-2.227
	L_{SL_1}	L_{SO}	10^{-5}	5.117×10^{-2}	5.429×10^{-2}	3.623	4.628×10^{-3}	9.700×10^{-1}	-2.227
	L_{SL_1}	L_{FD}	10^{-2}	4.107×10^{-2}	4.404×10^{-2}	3.382	4.268×10^{-3}	9.754×10^{-1}	-2.227
$\langle \lvert\nabla B_{\mathrm{LoS}}\rvert \rangle$	L_{MSE}	None	–	2.397	6.542×10^{-1}	3.360×10^{1}	1.064×10^{-2}	9.946×10^{-1}	0.000
	L_{SL_1}	L_{SO}	10^{-5}	1.302	3.597×10^{-1}	1.545×10^{1}	5.268×10^{-3}	9.983×10^{-1}	0.000
	L_{SL_1}	L_{FD}	10^{-2}	9.963×10^{-1}	2.757×10^{-1}	1.150×10^{1}	4.162×10^{-3}	9.992×10^{-1}	0.000

TABLE 4.7: Test–set temporal–evolution metrics for the derived physical parameters computed from the original and reconstructed LoS magnetograms. The table reports results for the selected models: the reconstruction–only baseline, the Sobel–regularized model, and the finite–difference gradient–regularized model. Lower values are better for *MAE*, *sMAE*, *DTW distance*, and *sDTW distance*, while higher values are better for ρ.

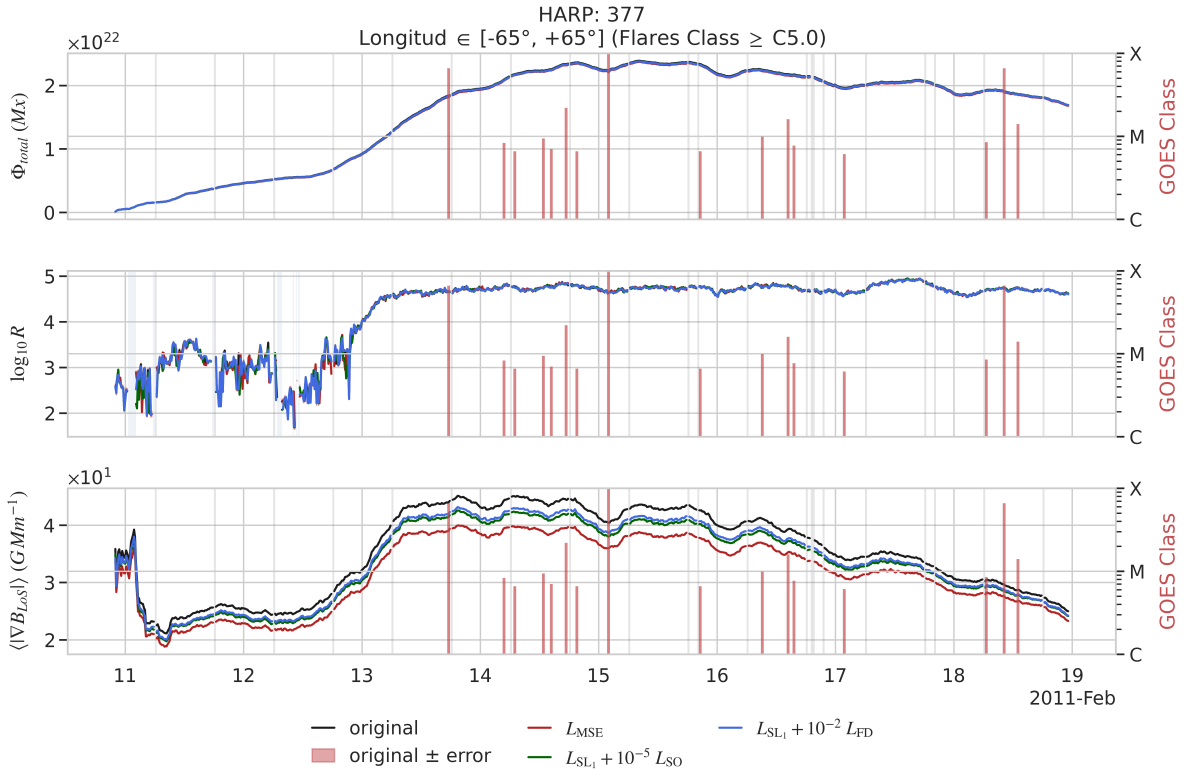


FIGURE 4.7: Temporal evolution of the derived physical parameters for HARP 377. The panels show the total unsigned LoS magnetic flux Φ, the gradient–weighted neutral line length R, and the mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$. Black curves indicate the parameters computed from the original magnetograms, while colored curves indicate the parameters computed from magnetograms reconstructed with the selected models. Vertical red lines indicate flares of class $\geq$ C5.0. Gray shaded regions indicate missing observations, while blue shaded regions indicate intervals with no PIL formation.

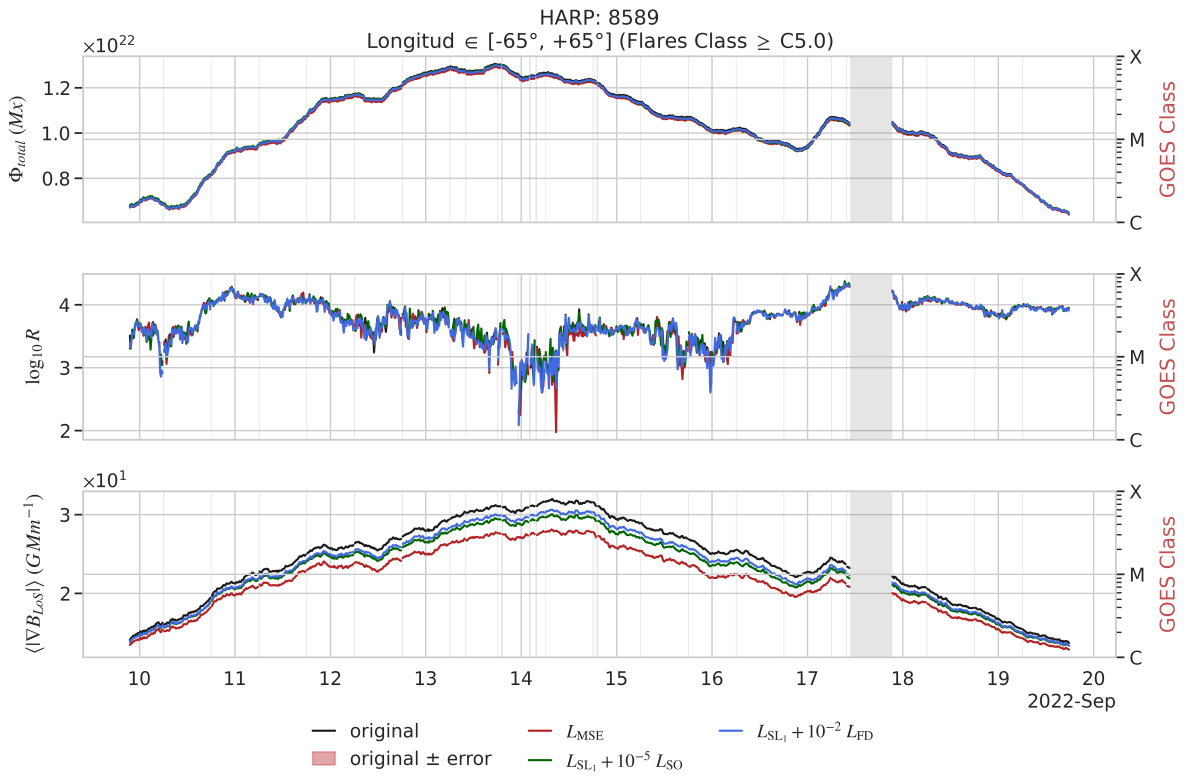


FIGURE 4.8: Temporal evolution of the derived physical parameters for HARP 8589. The panels show the total unsigned LoS magnetic flux Φ, the gradient–weighted neutral line length R, and the mean LoS magnetic–field gradient $\langle |\nabla B_{\mathrm{LoS}}| \rangle$. Black curves indicate the parameters computed from the original magnetograms, while colored curves indicate the parameters computed from magnetograms reconstructed with the selected models. Vertical red lines indicate flares of class $\geq$ C5.0. Gray shaded regions indicate missing observations, while blue shaded regions indicate intervals with no PIL formation.

4.4.4 Summary of Key Findings

The analysis conducted in this chapter was designed to answer the two questions formulated at the beginning of the results section. We now summarize the principal findings by addressing these questions in turn.

1. Do the selected training objectives improve the overall test–set image reconstruction quality?

The results show that the choice of training objective has a direct impact on reconstruction performance. Among the selected models, the best overall reconstruction performance is obtained with the finite–difference gradient–regularized model, $L_{\mathrm{SL}_1} + 10^{-2} L_{\mathrm{FD}}$. This configuration achieves the lowest test–set *MSE* (3.000×10^{-6}), the lowest *MAE* (9.160×10^{-4}), and the highest *MSSIM* (9.967×10^{-1}), indicating that combining a robust reconstruction loss with an explicit constraint on local gradients improves both pixel–wise accuracy and structural preservation.

The reconstruction–only baseline gives the lowest *KL divergence* (3.001×10^{-2}), suggesting better agreement between the global pixel–value distributions of the original and reconstructed magnetograms. This highlights a trade–off between preserving the global pixel–value distribution and preserving local spatial structure. While the reconstruction–only objective better preserves the global histogram, the gradient–regularized objectives better preserve magnetic morphology, local sharpness, and spectral power at high values of k.

The two representative reconstruction examples further support this interpretation. The selected models recover the large–scale polarity structure of both complex flare–productive and quieter ARs. However, the differences between the selected models become more visible near compact flux concentrations, strong magnetic–field gradients, and PILs. In these regions, the reconstruction–only baseline produces larger residuals, while the Sobel–regularized and finite–difference gradient–regularized models better preserve the local morphology.

2. Do improvements in image–level reconstruction translate into physically consistent temporal evolution of derived magnetic–field parameters?

The analysis of the temporal evolution of physical quantities derived from LoS magnetograms shows that the selected models behave differently depending on the parameter considered. The total unsigned LoS magnetic flux Φ is recovered with high accuracy from the reconstructed magnetograms and remains temporally coherent across all selected models, with correlations close to unity and zero *CCF peak lag*. This is consistent with the integrated nature of Φ, which makes it less sensitive to localized reconstruction errors.

The gradient–weighted neutral line length R is also recovered with good accuracy, with the finite–difference gradient–regularized model achieving the best performance. However, the *Pearson correlation coefficient* and *CCF peak lag* values depart more from their ideal values, and the standardized errors (*sMAE* and *sDTW distance*) are higher than in the case of Φ, indicating that the magnetic features on which the computation of R depends are harder to reconstruct. This behavior is expected because R depends on magnetic flux

near PILs, which are localized regions characterized by strong fields, sharp gradients, and high magnetic complexity.

The clearest benefit of gradient regularization is observed for the mean LoS magnetic–field gradient $\langle|\nabla B_{\mathrm{LoS}}|\rangle$, for which the finite–difference gradient–regularized model substantially reduces the standardized errors relative to the reconstruction–only baseline. This result is consistent with the definition of the parameter, which depends directly on local spatial derivatives of the magnetic field. Overall, the temporal diagnostics show that improvements observed at the level of individual magnetogram reconstructions translate into more physically consistent temporal evolution, especially when the derived quantities depend on local gradients.

4.5 Conclusions

In this chapter, we presented the first experimental phase of the thesis, focused on learning compact representations of active–region magnetic structure from LoS magnetograms. We trained a CAE to reconstruct magnetograms using different training objectives, and reconstruction quality was used as an indirect assessment of whether the learned latent representation preserves physically meaningful information about the spatial organization of the photospheric magnetic field. Since the architecture does not include skip connections or normalization layers, the reconstruction must rely only on the information encoded in the latent representation, making the comparison between training objectives particularly informative.

The results show that the finite–difference gradient–regularized objective, $L_{\mathrm{SL}_1} + 10^{-2} L_{\mathrm{FD}}$, provides the best compromise among the selected models. It improves pixel–wise reconstruction accuracy, structural similarity, preservation of spectral power at high values of k, and the temporal consistency of derived magnetic–field parameters. The reconstruction–only L_{MSE} baseline remains more favorable for reproducing the global pixel–value distribution, while the Sobel–regularized model generally shows intermediate behavior between the baseline and the finite–difference model.

These findings indicate that explicitly constraining local magnetic–field gradients during training helps the CAE preserve both large–scale magnetic organization and small–scale magnetic complexity. This is particularly important for physical diagnostics such as R and $\langle|\nabla B_{\mathrm{LoS}}|\rangle$, which depend on PIL structure and local magnetic–field gradients. The learned representation therefore retains information relevant not only to image reconstruction, but also to the temporally consistent recovery of derived physical quantities such as Φ, R, and $\langle|\nabla B_{\mathrm{LoS}}|\rangle$.

At the same time, the results suggest directions for further improvement. In particular, future work could experiment with adding additional terms to the reconstruction loss that explicitly promote overall preservation of structure in the Fourier domain. Further experiments could explore weighting different frequencies differently, or targeting specific spatial scales without compromising local magnetic morphology.

Overall, this chapter shows that the design of the training objective plays a central role in determining the physical consistency of the learned representation. The best–performing encoder preserves information about both global magnetic flux and localized magnetic complexity, supporting its potential use as a physically informed representation for downstream tasks.

Chapter 5

Phase II: Spatio–Temporal Forecasting via Sequence Encoding

"Begin at the beginning, . . . and go on till you come to the end: then stop."

Lewis Carroll,
Alice's Adventures in Wonderland

5.1 Introduction

In Chapter 4, we addressed the first phase of the flare–forecasting framework by learning compact representations of individual solar active–region magnetograms. In that phase, a convolutional autoencoder (CAE) was trained to reconstruct individual active–region line–of–sight (LoS) magnetograms, with the reconstruction task used as a proxy objective for learning a compact latent representation of the magnetic–field structure (Hinton and Salakhutdinov, 2006; Goodfellow et al., 2016). The encoder learned in this way therefore provides a compact spatial representation of the magnetic–field structure of an active region (AR) at a single time. However, flare productivity is not determined only by the magnetic configuration at one instant. The temporal evolution of the AR also contains important information, since the build–up, persistence, and reorganization of magnetic structures may precede the occurrence of major flaring activity (Wang and Liu, 2015; Toriumi and Wang, 2019).

The next natural step is therefore to move from representation learning to sequence–based flare forecasting. Instead of treating active–region magnetograms independently, they are organized into sequences that represent the temporal evolution of each AR. Each sequence spans an observational window and is assigned a label according to the flare activity occurring in a subsequent forecasting window. The forecasting task is then formulated as a binary classification problem in which the model predicts whether the AR will produce at least one M– or X–class flare during the forecasting window.

The proposed model combines the spatial representations learned in the previous phase with a temporal modeling component and a classification component. Rather than using the original magnetograms directly, this chapter exploits the previously trained encoder to map each observation into a compact latent representation. Each magnetogram in the input sequence is first passed through the frozen convolutional encoder trained in Chapter 4. Attention pooling is subsequently applied so that each magnetogram is represented as a single feature vector. The resulting sequence of feature vectors is processed by a long short–term memory (LSTM) network, which produces a representation of the temporal evolution used for binary classification. In this way, we separate the learning of spatial magnetic–field representations from the modeling of their temporal evolution.

The first aspect to consider is the definition of observational and forecasting windows. We evaluate four temporal configurations, with paired observational and forecasting windows of 24, 12, 6, and 3 h. Since major flares are rare, the resulting datasets are strongly imbalanced. To address this issue, we evaluate two random undersampling strategies: static undersampling, in which a fixed balanced subset of negative samples is used throughout training, and dynamic undersampling, in which the negative class is resampled at each epoch to expose the model to a wider variety of majority–class samples.

The objective of this chapter is therefore twofold. First, it evaluates whether sequence–based modelling provides effective forecasting performance for M– and X–class flares. Second, it examines how performance depends on the temporal configuration of the input sequence and on the undersampling strategy used during training. Model performance is assessed on the independent test set using classification metrics, threshold–independent discrimination metrics, probability–distribution analysis, and calibration diagnostics.

The chapter is organized as follows. Section 5.2 describes the construction of the sequence–based dataset, including the data source, active–region split, sequence extraction and labeling procedure, random undersampling, and preprocessing. Section 5.3 introduces the forecasting architecture and the evaluation tools. Section 5.4 presents the model–selection results, threshold analysis, independent test–set performance, discrimination and probability–separation analyses, calibration analysis, and summary of findings. Finally, Section 5.5 concludes the chapter by discussing the main findings and the role of temporal sequence encoding in the forecasting framework developed in this thesis.

5.2 Data

5.2.1 Data Sources and Dataset Construction

The dataset used in this chapter is constructed from the same observational data source introduced in Chapter 4. In particular, we use LoS magnetogram cutouts of solar ARs obtained from the Space–weather HMI Active Region Patches (SHARP) data product. The selection of ARs follows the

same criteria described in Section 4.2.2. We consider SHARP records from May 2010 to December 2022, covering most of solar cycle 24 and the early phase of solar cycle 25. Only SHARPs associated with exactly one National Oceanic and Atmospheric Administration (NOAA) active–region number are retained, allowing an unambiguous association with events listed in the Geostationary Operational Environmental Satellite (GOES) flare catalog. To mitigate near–limb projection effects, observations were restricted to ARs located within $\pm 65^{\circ}$ of the central meridian.

GOES flare events are associated with SHARP observations using the same time–alignment procedure described in Section 4.2.2. For each flare, the SHARP record closest to the flare peak time is selected within a ± 12–minute window. When no SHARP record is available within this tolerance interval, an empty placeholder record is created to preserve the information that a flare of a certain class occurred at that time. Preserving this information was useful for the subsequent active–region–level partitioning into training, validation, and test sets.

In the single–image reconstruction task considered in Chapter 4, missing and poor–quality records did not represent a problem, since magnetograms were processed independently rather than as part of a time–ordered sequence. Consequently, at training time, placeholder records were discarded, since no actual data existed at those times, with no further consideration. Poor–quality records were also discarded. The present chapter, however, addresses flare forecasting from time–ordered sequences of observations; therefore, missing and poor–quality records require a different treatment. The details are provided in Section 5.2.3.

5.2.2 Train/Validation/Test Split

The train, validation, and test partition used in this chapter follows the same active–region–level splitting strategy described in Section 4.2.3, in which ARs were first separated into two groups: those that produced at least one M– or X–class flare and those that did not. Each group was then split into training, validation, and test subsets using an approximate 70% / 15% / 15% target ratio. The groups were subsequently recombined, yielding an active–region–level stratified split that preserves a comparable fraction of M/X–flaring regions across subsets while maintaining independence at the active–region level. The split is applied before sequence extraction and labeling. Consequently, all sequences generated from a given AR are assigned to the same subset, preventing direct information leakage between the training, validation, and test sets.

5.2.3 Sequence Extraction and Labeling

Sequences are drawn from each active–region time series using a sliding–window approach similar to those adopted by Angryk et al. (2020), Sun et al. (2022), and Guastavino et al. (2022). Each sequence, also referred to as a sample, spans a T-hour window, with $T \in \{24, 12, 6, 3\}$, and the window

is shifted forward in 12 min increments to construct successive sequences. Thus, each new sample begins 12 min after the start of the previous one, matching the nominal temporal resolution of the SHARP data. Sequences are labeled according to the flare activity occurring in the subsequent T-hour window. The T-hour window used as input to the model is referred to as the observational window, while the subsequent T-hour window used for labeling is referred to as the forecasting window.

During the process of sequence extraction and labeling, a sequence is retained only if it satisfies one of the following criteria: (i) at least one M– or X–class flare occurs during the forecasting window, in which case the sequence is labeled as positive; or (ii) no flare of any class occurs during either the observational window or the forecasting window, in which case the sequence is labeled as negative (see Table 5.1 and Figure 5.1). This positive/negative labeling convention follows the common practice in solar flare forecasting, where M– and X–class flares are treated as the events of interest because they correspond to strong and potentially geoeffective events (Bobra and Couvidat, 2015; Barnes et al., 2016; Leka et al., 2019; Georgoulis et al., 2021).

	Forecasting Window		
Observational Window	Quiet	Weak	Strong
Quiet	−	?	+
Weak		?	+
Strong			+

TABLE 5.1: Definition of sequence class labels for the binary classification task. Positive samples are denoted by "+", and negative samples by "−". Blank entries correspond to a decay in flare activity, while "?" entries correspond to sequences leading to weak–flare activity; both cases are excluded under the labeling scheme adopted for the present task.

We note that two evolutionary types are excluded: (i) decaying flare activity, corresponding to sequences in which flare activity decreases from the observational window to the forecasting window, such as weak–to–quiet, strong–to–quiet, or strong–to–weak transitions; and (ii) sequences that produce only weak flares in the forecasting window, such as quiet–to–weak or weak–to–weak transitions (see Table 5.1 and Figure 5.1). These cases were deliberately excluded to define a controlled discrimination problem between sequences preceding major flare activity and sequences that remain completely flare–quiet over both the observational and forecasting intervals. Previous studies have adopted similar exclusions to reduce label noise and enhance class separation (Sun et al., 2022), while other studies have shown that heterogeneous evolutionary types may complicate the forecasting problem (Guastavino et al., 2022).

Under the broader definition used in all–clear flare forecasting, the quiet–to–quiet sequences retained as negative samples are themselves all–clear cases, since no event above the target threshold occurs during the forecasting window (Barnes et al., 2016). However, they represent only a restricted subset

of possible all–clear conditions. The labeling strategy adopted here should therefore be understood as a deliberate methodological choice rather than as an attempt to reproduce the full range of evolutionary states encountered in an operational forecasting setting. Its purpose is to reduce label heterogeneity and assess whether the temporal evolution of the learned magnetic representations contains information associated with major flare activity.

As already mentioned, the SHARP data series contain gaps and, in certain cases, records are missing precisely at the time of a flare, sometimes even for large flares. These cases must be treated carefully, since missing records can disrupt the sequential structure of the data and affect the label assigned to a sequence. To account for such situations, an empty placeholder record is created whenever a SHARP observation is unavailable. The placeholder record also contains the information of whether a flare of any magnitude occurred at that time, using the flare peak time as reference.

When constructing the observational and forecasting windows, missing records are handled differently depending on where they occur. If a placeholder record falls inside the observational window, the corresponding sequence is discarded, because the forecasting model requires a complete, fixed–length input sequence with regular temporal spacing. If a placeholder record falls inside the forecasting window, the sequence can instead be retained, because the forecasting window is used only to determine the label. In other words, placeholder records have the unique purpose of preserving the temporal cadence of the data and the information of whether a flare of a given magnitude occurred at that time, even though no corresponding magnetogram is available (see Figure 5.2). Sequences are also discarded if the observational window contains any poor–quality records, defined as records with `QUALITY` flags greater than or equal to 65536[1].

[1] `http://jsoc.stanford.edu/doc/data/hmi/Quality_Bits/QUALITY.txt`

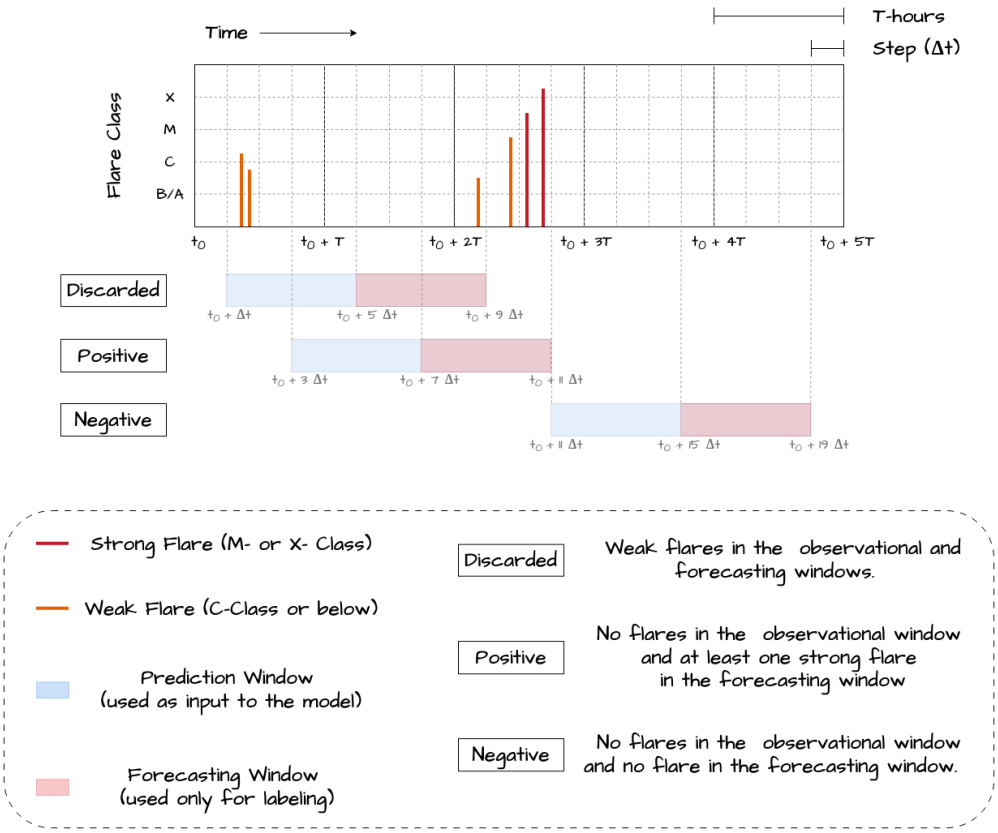


FIGURE 5.1: Example schematic representation of the sequence selection and labeling procedure. Each candidate sequence is defined by an observational window of duration T, sampled with cadence Δt, followed by a forecasting window of duration T used for label assignment. A sequence is labeled as positive if at least one M– or X–class flare occurs during the forecasting window, and as negative if no flare occurs during either the observational or forecasting window. Sequences corresponding to a decay in flare activity or to cases with only weak flares during the forecasting window are excluded from the classification task.

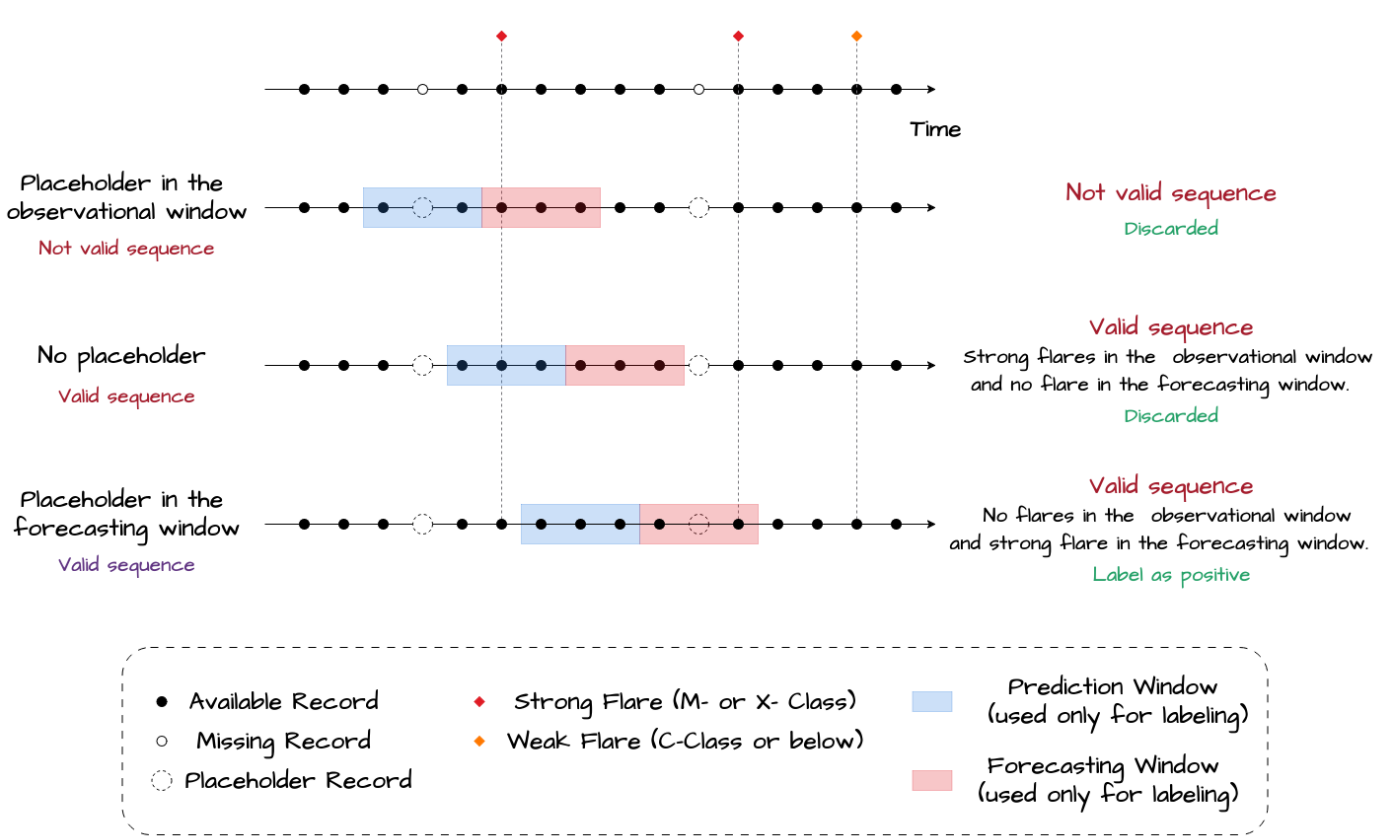


FIGURE 5.2: Example schematic representation of the treatment of missing records during sequence construction. When a record is missing, a placeholder is created. A candidate sequence is discarded when a placeholder record falls within the observational window, because the model input must be complete and regularly sampled. If a placeholder falls within the forecasting window of a candidate sequence with a complete observational window, the sequence can still be retained, since the forecasting window is used only for label assignment.

The number of samples generated by the above procedure for each time window is presented in Table 5.2. As shown, negative samples considerably outnumber positive samples. To address this class imbalance, the negative class is randomly undersampled to match the number of positive samples. Additional details on this process are provided in Section 5.2.4.

Observational Window	Forecasting Window	Sampling	Positive	Positive (%)	Negative	Negative (%)
24H	24H	36M	9,867	4.71	199,754	95.29
12H	12H	12M	8,494	2.33	355,974	97.67
6H	6H	12M	6,067	1.22	492,617	98.78
3H	3H	12M	3,713	0.62	592,115	99.38

TABLE 5.2: Number and percentage of positive and negative sequences generated for each temporal configuration. A temporal configuration is defined by the duration of the observational window in hours, the duration of the forecasting window in hours, and the sampling cadence of the input sequence in minutes.

For the 24 h samples, temporal pooling is applied during training, so the effective temporal cadence of the sequence is 36 min instead of 12 min. This is a purely architectural choice. At the original 12 min cadence, a 24 h sequence consists of 120 images of 256×512 pixels. Stored as 64 bit floating–point arrays during preprocessing and later converted to 32 bit floating–point arrays for GPU training, these sequences exceed the computational capacity available for this study.

5.2.4 Random Undersampling

Class imbalance refers to the situation in which one class is significantly overrepresented compared to others in a dataset. This is the case in flare forecasting, since most ARs either do not produce flares at all or produce only low–magnitude events (A–, B–, or C–class), resulting in a dataset heavily dominated by non–flaring and weakly flaring cases, with relatively few high–magnitude flares (e.g., M– or X–class). In a statistical analysis carried out by Georgoulis (2012) of a sample of 370 ARs observed with Solar and Heliospheric Observatory (SOHO)/Michelson Doppler Imager (MDI) during Solar Cycle 23, 77 regions, about 21%, produced at least one major flare, including 60 regions, about 16%, that produced at least one M–class flare and 17 regions, about 5%, that produced at least one X–class flare, while the remaining 293 regions, about 79%, were non–major–flaring. According to the multivariate time–series dataset compiled by Angryk et al. (2020) from 4,098 Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) SHARP active–region data observed between May 2010 and December 2018, corresponding to Solar Cycle 24, only 27, about 0.7%, contained

X–class flares and 178, about 4.3%, contained M–class flares, while 3,293, about 80%, did not contain any flares, including B– or C–class events.

Class imbalance is a well–documented problem in machine learning, because models trained on imbalanced dataset tend to be biased toward the overrepresented, or majority, class, leading to poor performance on the underrepresented, or minority, class (He and Garcia, 2009; He and Ma, 2013; Johnson and Khoshgoftaar, 2019; Ahmadzadeh et al., 2021). Class imbalance also makes model evaluation in flare forecasting challenging, since standard performance metrics can be strongly affected by the underlying class distribution (Woodcock, 1976; Jolliffe and Stephenson, 2012). The issue is even more significant in operational settings, where missing rare but potentially geoeffective events may be more consequential than misclassifying more common quiet periods or weakly flaring cases (Barnes et al., 2016; Leka et al., 2019; Georgoulis et al., 2021). More generally, this is a characteristic issue of rare–event prediction, where events of the greatest practical relevance are, by definition, rare.

To mitigate class imbalance, data–level strategies such as undersampling the majority class and oversampling the minority class are commonly employed (He and Garcia, 2009; He and Ma, 2013). Undersampling reduces the dominance of majority–class examples by randomly selecting a subset for training; however, because it discards data, it may also remove informative examples. Oversampling, on the other hand, increases the representation of rare events either by duplicating existing minority–class samples or by generating synthetic ones, as in the synthetic minority over–sampling technique (SMOTE; Chawla et al., 2002), where new samples are produced by interpolating between neighboring minority–class instances. While oversampling preserves all available data, duplicated samples may increase the risk of overfitting, and synthetic samples may be unrealistic, thereby adding noise. Nevertheless, training on a more balanced dataset encourages the model to learn better decision boundaries for all classes, rather than being biased toward the majority class, often improving the detection of minority–class instances.

Data augmentation is another technique used to enhance the diversity and quantity of training data. It typically involves applying label–preserving transformations to existing samples; for image data, these transformations may include rotations, flips, crops, or scaling operations (Shorten and Khoshgoftaar, 2019). In some cases, more advanced approaches employ generative models, such as variational autoencoders (VAEs; Kingma and Welling, 2013) or generative adversarial networks (GANs; Goodfellow et al., 2014), to synthesize entirely new, realistic examples of minority–class instances.

In addition to data–level strategies, algorithm–level techniques can mitigate class imbalance without altering the underlying dataset (He and Garcia, 2009; He and Ma, 2013). These methods act directly on the learning algorithm, most commonly by modifying the training objective so that errors on minority–class receive greater emphasis than errors on majority–class samples. A widely used approach is cost–sensitive learning, in which class–dependent penalties are incorporated into the loss function, for example via

weighted cross–entropy (Khan et al., 2017). Related objective–level reweighting methods, such as focal loss, address imbalance by dynamically down–weighting well–classified samples and focusing learning on harder cases (Lin et al., 2017).

In this work, we address class imbalance by employing two strategies based on random undersampling. Specifically, samples from the negative class are randomly selected so that the numbers of positive and negative samples are equal. The first strategy, hereafter referred to as static undersampling, consists of performing this operation once and using the selected subset throughout training. This approach simplifies the training process and reduces computational overhead while still alleviating the dominance of the majority class. The second strategy, hereafter referred to as dynamic undersampling, consists of selecting a new random subset of the negative class at each training epoch. This method introduces greater variability in the negative samples seen by the model over time, promoting better generalization and reducing the risk of overfitting to a fixed subset. To ensure a fair comparison between the two strategies, models were trained using either static or dynamic undersampling, while validation and testing were performed on the same fixed balanced sets. The validation and test sets are therefore balanced by construction, with equal numbers of positive and negative samples.

5.2.5 Preprocessing

The preprocessing steps are essentially the same as those described in Section 4.2.4. The main difference is that the operations previously applied to individual LoS magnetograms are now applied consistently to all magnetograms in a sequence. This ensures that all magnetograms within the same sequence undergo a consistent spatial transformations, avoiding the introduction of artificial frame–to–frame variations.

As in the previous phase, magnetograms are standardized to a fixed input size of 256×512 pixels. Larger active–region cutouts are randomly cropped to this target size, while smaller cutouts are padded using an estimate of the background magnetic field. The background field used for padding is approximated as the mean of the pixel values within the IQR, as described in Section 4.2.4. In the present sequence–based setting, however, both cropping and padding are defined once at the sequence level rather than independently for each frame.

Pixel values are scaled using the global minimum and maximum values computed from the training set, which map the training–set range to $[-1, 1]$. The same scaling parameters are then applied to the validation and test sets. As discussed in Section 4.2.4, a known consequence of this fixed training–derived transformation is that held–out values may be mapped outside the nominal $[-1, 1]$ interval. Missing values are replaced with the median of the valid pixels in each frame.

5.3 Methodology

5.3.1 Model

The classifier presented in this chapter builds on the model trained in Chapter 4. In that phase, a CAE was trained to learn low–dimensional spatial representations of individual active–region LoS magnetograms through an image reconstruction objective. In the present phase, the encoder learned as part of the autoencoder is used as a spatial feature extractor. Each magnetogram in an input sequence is mapped to a low–dimensional latent feature vector, and the resulting sequence of learned representations is processed by a recurrent classifier to predict whether the AR will produce an M– or X–class flare during the forecasting window.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_T}\}$ denote an input sequence of preprocessed LoS magnetograms, where each frame $\mathbf{x}_t \in \mathbb{R}^{1\times H\times W}$ corresponds to one time step in the observational window. Each magnetogram is first passed through a convolutional encoder $E(\cdot)$, producing a latent feature tensor,

$$\mathbf{z}_t = E(\mathbf{x}_t), \tag{5.1}$$

so that $\mathbf{z}_t \in \mathbb{R}^{512\times H/32\times W/32}$. The encoded tensor $\mathbf{z}_t$ is subsequently reduced to a feature vector $\mathbf{f}_t$,

$$\mathbf{f}_t \in \mathbb{R}^{512}. \tag{5.2}$$

This spatial aggregation is commonly performed by averaging each feature map. Here, however, a non–local self–attention block is applied before the average–pooling operation (Wang et al., 2018b). This block allows spatial locations within $\mathbf{z}_t$ to interact before the feature tensor is aggregated. Since the attention block is trained together with the forecasting classifier, it can learn to emphasize encoded spatial structures that are more relevant for predicting M– or X–class flares.

Applying this procedure independently to all frames in the observational window produces a sequence of feature vectors,

$$\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{N_T}\}. \tag{5.3}$$

The temporal evolution of these feature vectors is modeled using a LSTM layer. The LSTM receives $\mathbf{F}$ as input and produces a hidden state at each time step,

$$\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{N_T}\} = \mathrm{LSTM}(\mathbf{F}). \tag{5.4}$$

Only the final hidden state is retained, since it provides a compact representation of the temporal information captured by the LSTM over the full observational window,

$$\mathbf{r} = \mathbf{h}_{N_T}. \tag{5.5}$$

The sequence representation $\mathbf{r}$ is then passed to a fully connected classification head, which produces a scalar logit ℓ. This logit is associated with the probability that the input sequence precedes at least one M– or X–class flare during the forecasting window. The probability of the positive class is therefore obtained as

$$\hat{p} = \sigma(\ell), \tag{5.6}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

The spatial module, shown as the blue block in Figure 5.3, consists of a convolutional encoder followed by an attention–pooling stage. The convolutional encoder has the same structure as the one used in the previous phase: five convolutional blocks, each consisting of a two–dimensional convolution followed by a Rectified linear units (ReLU) activation. The convolutional layers use kernel size 3×3, stride 2, and padding 1, progressively reducing the spatial resolution while increasing the number of feature channels. The attention block uses group normalization, a Swish activation, and 1×1 convolutions to compute query, key, and value projections. It is followed by an adaptive average pooling layer and dropout with probability $p = 0.1$. The temporal module, shown as the green block, is implemented as a single–layer LSTM followed by dropout with probability $p = 0.1$. The classification module, shown as the red block, consists of a linear layer, a ReLU activation, dropout with probability $p = 0.1$, and a final linear layer with one output unit.

The encoder is frozen during this phase, meaning that its parameters are kept fixed throughout training. The encoder used here corresponds to the autoencoder trained with the L_{MSE} reconstruction objective. The model is trained using a binary cross–entropy with logits objective, which combines the sigmoid activation and binary cross–entropy loss in a numerically stable form (Bishop and Nasrabadi, 2006; PyTorch Contributors, n.d.). Optimization is performed using the Adam optimizer with its default momentum parameters. Models are trained for up to 100 epochs using a batch size of 32. Early stopping is applied based on the validation loss, with a patience of 5 epochs and a minimum improvement threshold of 10^{-3}. A model checkpoint is saved when the validation loss falls below 0.35 and is subsequently updated whenever a lower validation loss is achieved.

The reconstruction–only encoder was retained as the baseline representation for the forecasting experiments, which focused on the temporal modeling of the encoded magnetogram sequences rather than on comparing the downstream performance of alternative representation–learning objectives. No downstream comparison between the reconstruction–only encoder and the gradient–regularized encoders evaluated in Chapter 4 was performed.

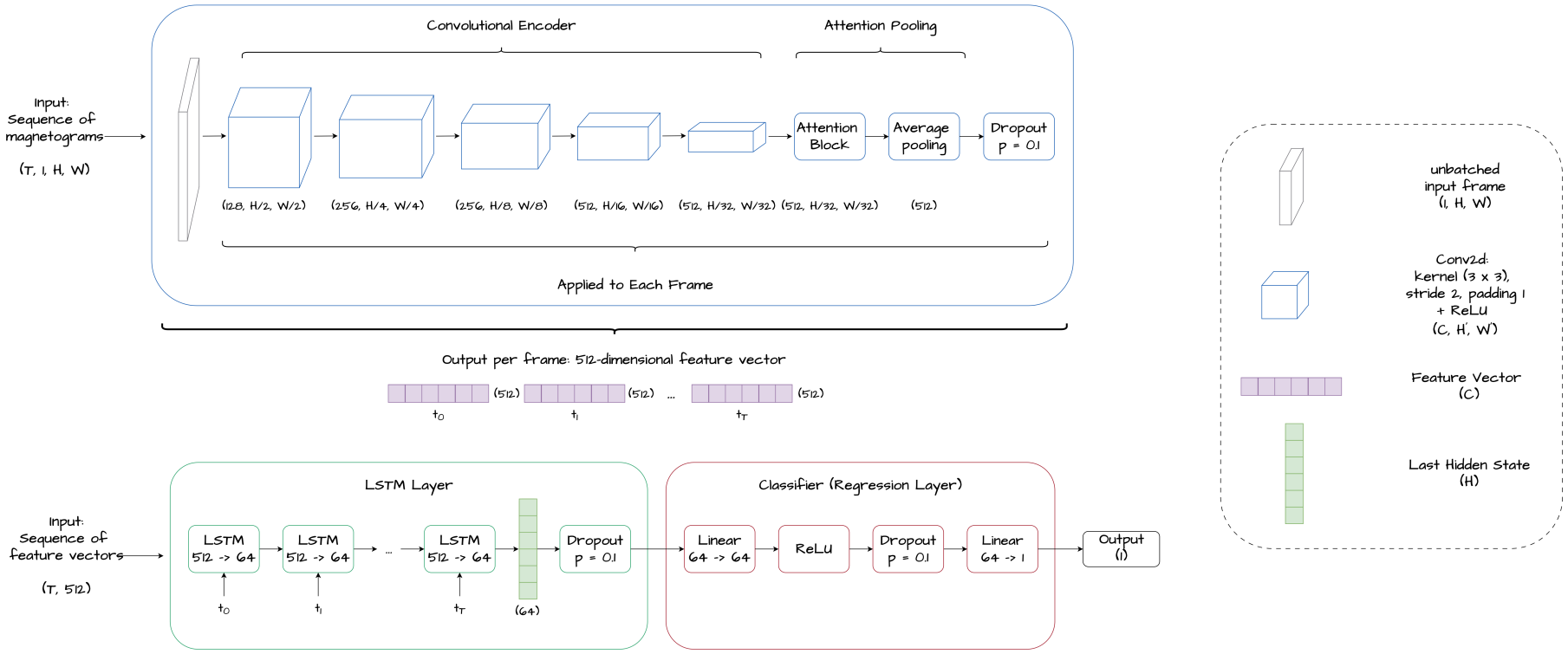


FIGURE 5.3: Schematic representation of the forecasting model architecture. Each magnetogram in the input sequence is processed independently by a shared frozen convolutional encoder with attention pooling, producing a 512–dimensional feature vector for each time step. The resulting sequence of feature vectors is passed to an LSTM layer, and the final hidden state is used by a fully connected classifier to produce the binary prediction.

The LSTM hidden size and the learning rate are treated as hyperparameters. In particular, we test hidden sizes of 32, 64, 128, and 256, and learning rates of 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. The final model configuration is selected on the basis of validation–set performance, as discussed in Section 5.4.1.

5.3.2 Evaluation Tools

The probability $\hat{p}$ of the positive class, defined in Equation 5.6, must be transformed into a binary label. To do so, a threshold (τ) must be chosen, with 0.5 representing a natural first choice. A sequence is classified as positive by the model when $\hat{p} \geq \tau$, and as negative otherwise. The performance of the proposed solar flare forecasting model is then evaluated using standard classification metrics derived from the confusion matrix. The confusion matrix summarizes the model's predictions in terms of four outcomes: *true positives (TP)* and *true negatives (TN)*, which represent correct predictions of flare occurrence and non–occurrence, respectively; and *false positives (FP)* and *false negatives (FN)*, which correspond to incorrect predictions. It provides a structured basis for analyzing the model's performance, from which several metrics are computed.

Accuracy (Equation 5.7) measures the proportion of correctly classified instances across the entire dataset. Although it is widely used, *Accuracy* can be misleading in the presence of class imbalance, where one class significantly dominates the other.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.7}$$

Precision (Equation 5.8) quantifies the proportion of predicted positive instances (flare forecasts) that are actually correct. It reflects the reliability of positive forecasts and is particularly useful in imbalanced datasets. However, it does not account for false negatives, which means it may overlook missed flare occurrences.

$$Precision = \frac{TP}{TP + FP} \tag{5.8}$$

Recall, also known as *Sensitivity* or the *true–positive rate (TPR;* Equation 5.9), evaluates the model's ability to detect actual flare occurrences, thereby reducing missed events. High *Recall* is particularly important in space weather forecasting, where undetected flares could have serious consequences.

$$Recall = \frac{TP}{TP + FN} \tag{5.9}$$

Based on the previous metrics, the *F1–score* (Equation 5.10) is defined and used to evaluate the balance between *Precision* and *Recall*. It is particularly useful for imbalanced datasets, as it provides a single score that accounts for both false positives and false negatives. The *F1–score* is computed as the harmonic mean of *Precision* and *Recall*, thereby ensuring that both the ability to detect flares and the reliability of predictions are considered. A high

F1–score indicates a well–balanced model that minimizes both missed flare events (*FN*) and false alarms (*FP*).

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{5.10}$$

Specificity, also known as the *true–negative rate (TNR*; Equation 5.11), is complementary to *Recall*, as it quantifies the model's ability to correctly identify non–flaring cases. A high *Specificity* indicates that the model rarely misclassifies quiet cases as flaring, thereby producing few false alarms.

$$Specificity = \frac{TN}{TN + FP} \tag{5.11}$$

Another widely used metric, the *true skill statistic (TSS*; Equation 5.12), builds upon these metrics and measures the model's ability to discriminate between flare and non–flare cases. It provides a balanced assessment that is less affected by class imbalance than *Accuracy*. A *TSS* value close to 1 indicates nearly perfect classification, while a value near 0 suggests performance no better than random guessing. It can be defined in terms of *Sensitivity* and *Specificity* as follows:

$$TSS = Sensitivity + Specificity - 1. \tag{5.12}$$

The last two metrics we introduce are the *false positive rate (FPR)*, defined as

$$FPR = 1 - Specificity, \tag{5.13}$$

and the *false negative rate (FNR)*, defined as

$$FNR = 1 - Recall. \tag{5.14}$$

The receiver operating characteristic (ROC) curve (Fawcett, 2006) illustrates the model's ability to discriminate between flare and non–flare cases by plotting the *TPR* (*Sensitivity*) against the *FPR* across different classification thresholds. A random classifier would follow the diagonal line $y = x$, whereas a high–performing model yields a curve that rises sharply toward the upper–left corner. The resulting *area under the receiver operating characteristic curve (ROC–AUC)* provides a comprehensive evaluation of model performance by analyzing the trade–off between the *TPR* and the *FPR*, with higher values indicating stronger discriminative capability.

The Precision–Recall (PR) curve (Davis and Goadrich, 2006) provides a complementary threshold–independent view of model performance. For this curve, *Recall* is plotted on the horizontal axis and *Precision* on the vertical axis. A trade–off usually exists between these two quantities; increasing *Recall* often means classifying more samples as positive, which may reduce *Precision* by increasing the number of false positives. The *area under the Precision–Recall curve (PR–AUC)* summarizes this relation, with higher values indicating that the model maintains high *Precision* over a broad range of *Recall* values.

Depending on the application context, different metrics may be prioritized. If false positives are considered tolerable, *Recall* becomes the most relevant metric. High *Recall* ensures the detection of most actual flares, reducing the risk of overlooking significant events. Conversely, in situations where false positives carry substantial costs, such as operational disruptions caused by false alarms, *Precision* becomes more critical. High *Precision* ensures that most predicted flares are actual flares, minimizing false alerts. However, the confusion matrix and the metrics derived from it depend on the chosen threshold; consequently, a threshold analysis is usually performed instead of adopting the naive value of 0.5. This analysis is used to determine the threshold that optimizes the metric of interest for the task under consideration. The threshold analysis is performed using predictions on the validation set, and the selected threshold is then used to compute the aforementioned metrics on the test set. In this work, *TSS* is selected as the metric to optimize, since it is a common choice in the flare–forecasting literature (e.g., Bobra and Couvidat, 2015; Liu et al., 2017a; Jonas et al., 2018; Sun et al., 2022; Guastavino et al., 2022).

Model performance is also analyzed using diagnostic plots. ROC curves summarize the trade–off between the *TPR* and *FPR* across thresholds. Predicted–probability distributions are used to assess the separation between positive and negative samples, while calibration curves compare predicted probabilities with observed flare frequencies.

The *Overlap* between the positive and negative predicted–probability distributions can be quantified as

$$Overlap = \sum_{m=1}^{M} \min\left(w_m^+, w_m^-\right), \tag{5.15}$$

where

$$w_m^+ = \frac{|B_m^+|}{N^+}, \qquad w_m^- = \frac{|B_m^-|}{N^-}. \tag{5.16}$$

Here, M is the number of probability bins, $|B_m^+|$ and $|B_m^-|$ are the numbers of positive and negative samples in bin B_m, respectively, and N^+ and N^- are the total numbers of positive and negative samples. The *Overlap* ranges from 0 to 1, with lower values indicating better separation between the two predicted–probability distributions.

In addition to evaluating how well the model discriminates between flare–positive and flare–negative sequences, we assess whether its predicted probabilities are calibrated. Calibration measures whether the predicted probabilities are consistent with the observed event frequencies, and therefore whether they can be interpreted as reliable probability estimates. The *Brier score* (Equation 5.17; Brier et al., 1950) measures the mean squared difference between the predicted probability $\hat{p}_i$ and the binary target y_i, where $y_i = 1$ for flare–positive sequences and $y_i = 0$ otherwise:

$$Brier\ score = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2; \tag{5.17}$$

As shown in Equation 5.17, confident incorrect predictions are penalized more strongly by this metric, so lower *Brier score* values indicate better probability estimates.

Calibration is also assessed by grouping the predicted probabilities into M probability bins. For each bin B_m, the mean predicted probability is compared with the observed positive fraction. The *expected calibration error (ECE*; Equation 5.18; Guo et al., 2017) summarizes the average calibration gap across bins, weighted by the number of samples in each bin:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| freq(B_m) - conf(B_m) \right|, \tag{5.18}$$

where N is the total number of samples, $|B_m|$ is the number of samples in bin B_m, $conf(B_m)$ is the mean predicted probability in bin B_m, and $freq(B_m)$ is the observed fraction of positive samples in that bin. A lower *ECE* indicates that, on average, predicted probabilities better match the empirical flare frequencies. The *maximum calibration error (MCE*; Guo et al., 2017; Equation 5.19) reports the largest calibration gap across all bins:

$$MCE = \max_{m} \left| freq(B_m) - conf(B_m) \right|. \tag{5.19}$$

While *ECE* measures the average calibration error, *MCE* identifies the largest localized deviation from perfect calibration. Therefore, a model may have a low *ECE* but a relatively high *MCE* if most probability bins are well calibrated but one bin shows a large discrepancy.

5.4 Results

This section presents the empirical evaluation of the proposed sequence–based forecasting framework for the prediction of M– and X–class solar flares. We compare the performance of sequence–based classifiers trained on four different temporal configurations, each defined by a specific observational window, temporal sampling cadence, and forecasting window. In addition, we evaluate two strategies for addressing the class imbalance inherent to the flare–forecasting problem: a static undersampling strategy, in which the negative training subset is fixed during training, and a dynamic undersampling strategy, in which the negative subset is resampled across training epochs.

The analysis is guided by three main questions: (1) which temporal configuration provides the strongest forecasting performance for major flare prediction? (2) how does dynamic undersampling compare with static undersampling across the evaluated temporal configurations? and (3) do the predicted probabilities provide meaningful class separation and reliable probabilistic information beyond threshold–dependent classification metrics?

To answer these questions, we follow a controlled evaluation protocol. First, the best model configuration for each sequence configuration and undersampling strategy is selected using validation–set loss. Second, the decision threshold of each selected model is optimized on the validation set using the *TSS*. Finally, the performance of the selected models is evaluated on the independent test set using threshold–dependent metrics computed with thresholds optimized on the validation set. These results are complemented by threshold–independent analyses of discrimination, probability separation between positive and negative classes, and calibration.

5.4.1 Model Selection

As described in Section 5.2.3, four sequence configurations were evaluated: *24H24H36M*, *12H12H12M*, *6H6H12M*, and *3H3H12M*. This naming convention indicates, in order, the duration of the observational window, the duration of the forecasting window, and the cadence of the input sequence. For example, *24H24H36M* denotes a 24 h observational window followed by a 24 h forecasting window, with input frames sampled every 36 min. Similarly, *6H6H12M* denotes a 6 h observational window, a 6 h forecasting window, and a 12 min input cadence.

We also evaluated two undersampling strategies: static undersampling, in which the same training subset was used throughout training, and dynamic undersampling, in which the majority–class samples were resampled during training at each epoch, as described in Section 5.2.4. The four sequence configurations and two undersampling strategies resulted in a total of 8 experimental settings.

The first step of the experimental analysis was to select the best model configuration for each sequence configuration and undersampling strategy. In this work, the LSTM *hidden size (HS)* and the learning rate (η) were treated as hyperparameters. A grid search was performed over four hidden sizes, 32, 64, 128, and 256, and four learning rates, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. Therefore, for each experimental setting, 16 hyperparameter combinations were trained and evaluated, giving a total of 128 runs.

For each experimental setting, the final model configuration was selected according to the minimum validation loss reached during training. Validation loss was used as the selection criterion instead of classification metrics such as *TSS* or *Accuracy* because these metrics depend on the choice of the decision threshold. At this stage, the goal was therefore to select the model that minimized a threshold–independent loss on the validation set. Threshold–dependent performance is analyzed separately in the following subsections.

Figure 5.4 and Figure 5.5 show the validation loss (upper row) and *TSS* (lower row) for static and dynamic undersampling, respectively. Each cell corresponds to a learning–rate and LSTM–hidden–size combination evaluated at the epoch at which the lowest validation loss was achieved. *TSS* is computed at a fixed threshold, $\tau = 0.5$, and is reported only for reference.

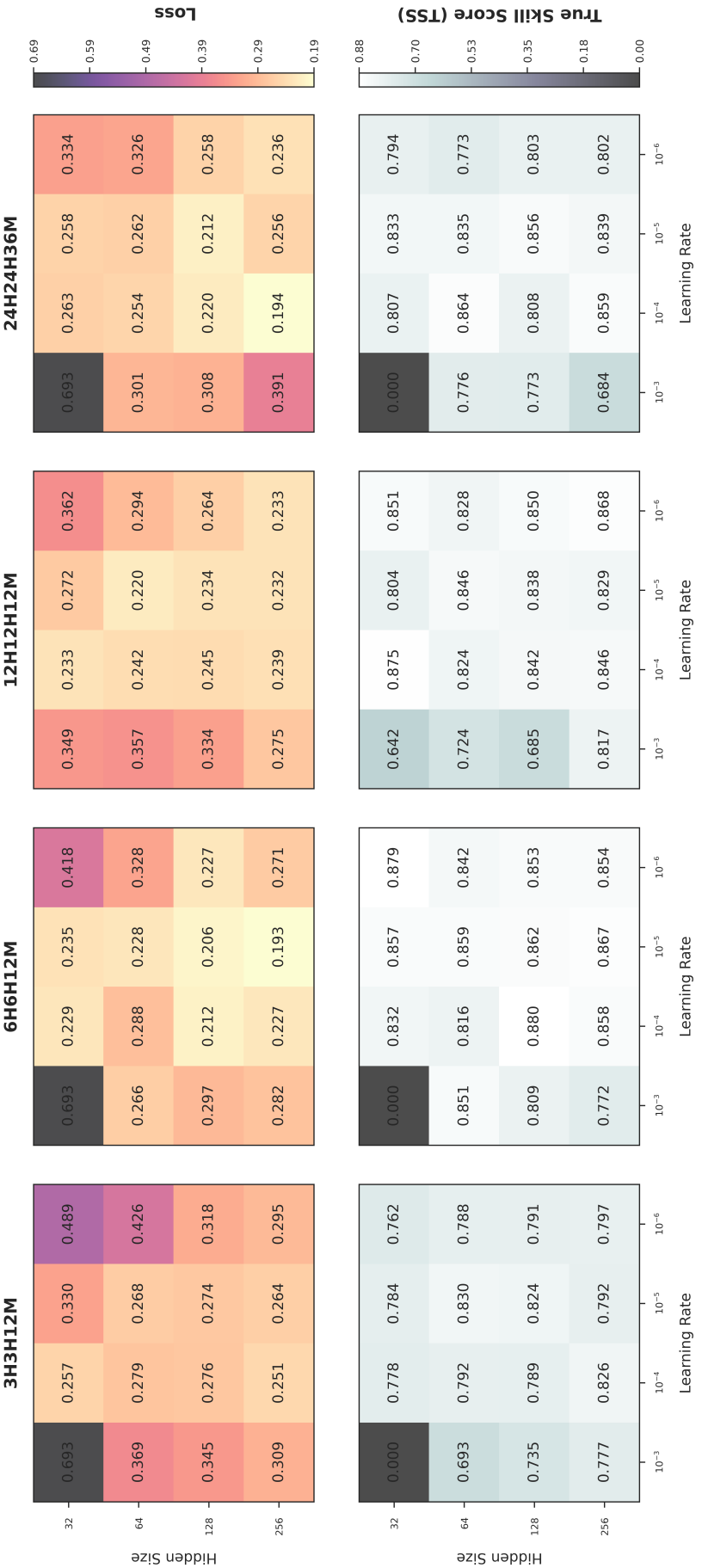


FIGURE 5.4: Validation–set model selection for the static–undersampling strategy. Each column corresponds to one temporal configuration, and each cell represents a combination of LSTM hidden size and learning rate. The upper row reports the minimum validation loss reached by each configuration, while the lower row reports the corresponding validation *TSS* evaluated at the same epoch using the fixed threshold $\tau = 0.5$.

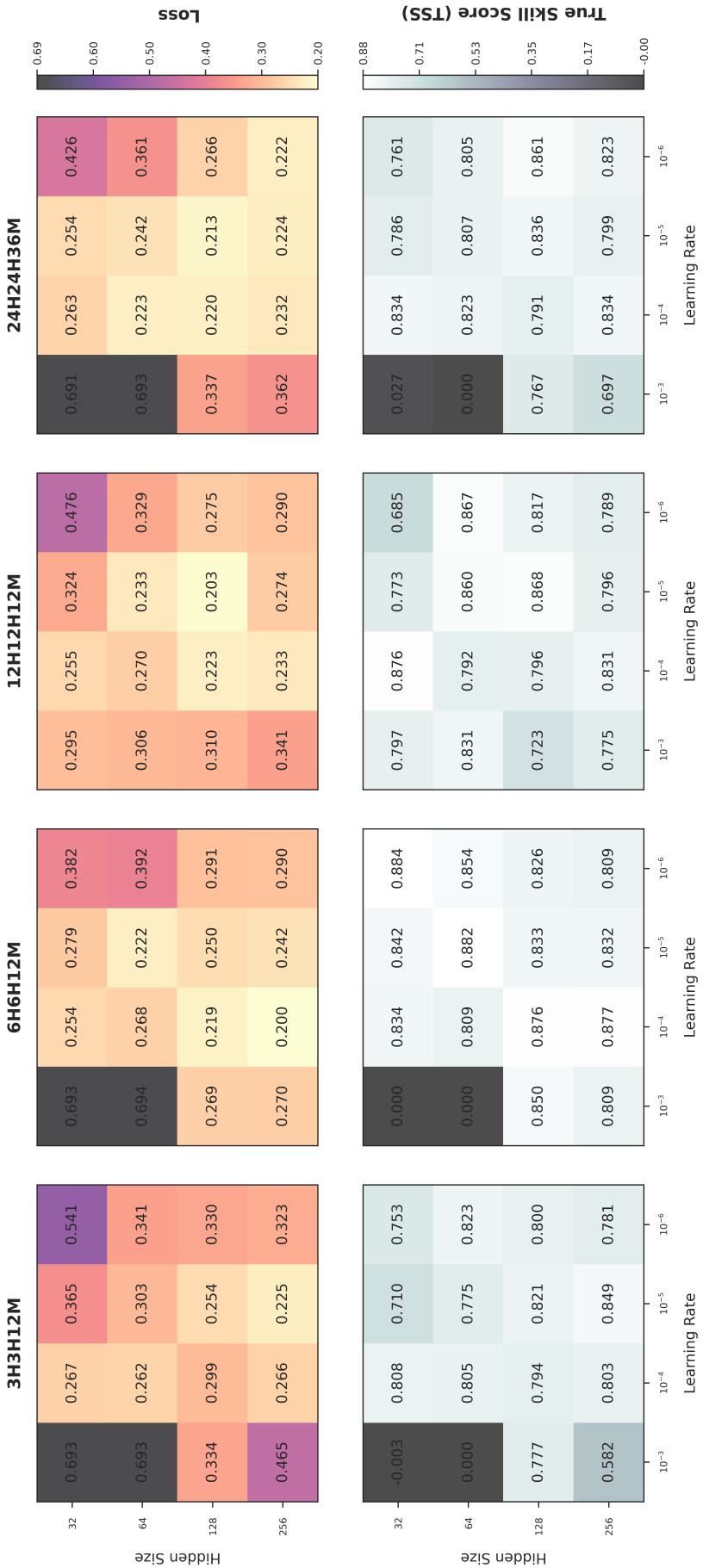


FIGURE 5.5: Validation–set model selection for the dynamic–undersampling strategy. Each column corresponds to one temporal configuration, and each cell represents a combination of LSTM hidden size and learning rate. The upper row reports the minimum validation loss reached by each configuration, while the lower row reports the corresponding validation *TSS* evaluated at the same epoch using the fixed threshold $\tau = 0.5$.

The results show that the learning rate had a stronger effect than the hidden size. In particular, the largest learning rate, 10^{-3}, often led to poor validation performance, with losses close to the uninformative regime and *TSS* values close to zero for some configurations. In contrast, intermediate learning rates, especially 10^{-4} and 10^{-5}, consistently produced lower validation losses and higher *TSS* values across the different sequence configurations. The presence of neighboring hyperparameter combinations with similar validation performance suggests that those selected are not isolated optima, but belong to broader regions of good performance in the hyperparameter space. For static undersampling, the minimum–validation–loss criterion selected $(HS = 256, \eta = 10^{-4})$ for *24H24H36M*, $(HS = 64, \eta = 10^{-5})$ for *12H12H12M*, $(HS = 256, \eta = 10^{-5})$ for *6H6H12M*, and $(HS = 256, \eta = 10^{-4})$ for *3H3H12M*. For dynamic undersampling, the minimum–validation–loss criterion selected $(HS = 128, \eta = 10^{-5})$ for *24H24H36M*, $(HS = 128, \eta = 10^{-5})$ for *12H12H12M*, $(HS = 256, \eta = 10^{-4})$ for *6H6H12M*, and $(HS = 256, \eta = 10^{-5})$ for *3H3H12M*.

Table 5.3 summarizes the model configurations selected from the hyperparameter search for each sequence configuration and undersampling strategy. The table also reports the best epoch, defined as the epoch at which the minimum validation loss was achieved, together with validation–set performance metrics computed at the fixed threshold $\tau = 0.5$ as a preliminary diagnostic. Across the selected model configurations, the lowest validation loss was obtained by the static *6H6H12M* configuration, with a validation loss of 0.193.

In general, dynamic undersampling produced a higher validation *TSS* than static undersampling for three of the four sequence configurations, namely *12H12H12M*, *6H6H12M*, and *3H3H12M*. These preliminary results suggest that dynamic undersampling generally improved validation skill, although the advantage was not uniform across all temporal configurations.

The selected model configurations also show differences in the balance between *Recall* and *Specificity*. Higher *Recall* is consistently associated with lower *Specificity*, indicating a stronger tendency to classify samples as flare–productive. This behavior is more evident for the static–undersampling models, particularly for the shorter *3H3H12M* temporal configuration, which achieved the highest *Recall*, 0.970, but the lowest *Specificity*, 0.856. In contrast, the dynamic–undersampling models generally show a more balanced relation between *Recall* and *Specificity*.

Although the validation–set metrics provide preliminary diagnostic information, at this stage the validation results only determine which hyperparameter configuration should be retained for each of the four temporal configurations and two undersampling strategies. A comparison between models requires evaluation on the independent test set, after fixing both the model configuration and the decision threshold. The validation analysis therefore serves only to define the eight candidate models carried forward to the subsequent threshold–optimization and test–set evaluation stages.

Undersampling Strategy	Temporal Configuration	Hidden State	Learning Rate	Epoch	Loss	TSS	Acc.	Prec.	Rec.	Spec.	F1	FPR	FNR
Static	24H24H36M	256	10^{-4}	6	0.194	0.859	0.930	0.948	0.909	0.950	0.928	0.050	0.091
	12H12H12M	64	10^{-5}	15	0.220	0.846	0.923	0.901	0.951	0.895	0.925	0.105	0.049
	6H6H12M	256	10^{-5}	8	0.193	0.867	0.933	0.912	0.959	0.908	0.935	0.092	0.041
	3H3H12M	256	10^{-4}	4	0.251	0.826	0.913	0.871	0.970	0.856	0.918	0.144	0.030
Dynamic	24H24H36M	128	10^{-5}	7	0.213	0.836	0.918	0.923	0.913	0.923	0.918	0.077	0.088
	12H12H12M	128	10^{-5}	8	0.203	0.869	0.934	0.920	0.951	0.918	0.935	0.083	0.049
	6H6H12M	256	10^{-4}	11	0.200	0.877	0.938	0.924	0.955	0.922	0.939	0.078	0.045
	3H3H12M	256	10^{-5}	15	0.225	0.849	0.925	0.911	0.941	0.908	0.926	0.092	0.059

TABLE 5.3: Best models for each temporal configuration and undersampling strategy. The selected model is the one achieving the minimum validation loss. Other classification metrics on the validation set are reported using the fixed threshold $\tau = 0.5$.

5.4.2 Threshold Analysis

The model does not directly predict a binary class label, such as "flare" or "no flare". Instead, it outputs a logit, ℓ, which is converted into the positive–class probability $\hat{p}$ through Equation 5.6. For example, a value $\hat{p} = 0.67$ indicates that the sequence is assigned a higher probability of being flare–productive than non–flare–productive. To convert this continuous score into a class label, we apply a decision threshold (τ):

$$\hat{p} \geq \tau \rightarrow \text{flare}, \qquad \hat{p} < \tau \rightarrow \text{no flare}. \tag{5.20}$$

The default choice is usually $\tau = 0.5$, but this value does not necessarily provide the best operating point for the forecasting problem. The decision threshold controls the balance between the *TPR* and *FPR*. A lower threshold makes the model more sensitive, increasing the number of predicted flares, but also increasing the number of false alarms. Conversely, a higher threshold makes the model more conservative, reducing false alarms but potentially increasing the number of missed detections. In our case, the threshold was optimized by maximizing *TSS* (Equation 5.12), which combines *Recall* and *Specificity* and rewards both the detection of flare cases and the rejection of non–flaring cases.

After selecting the best hyperparameter combination for each temporal configuration and undersampling strategy, the decision threshold was optimized on the validation set. The optimization was performed by evaluating *TSS* over a uniform threshold grid between 0 and 1. In cases where multiple thresholds achieved the same maximum *TSS*, the first threshold in the grid was selected, corresponding to the lowest threshold among tied optima. All thresholds were optimized only on validation data, and the resulting values were subsequently fixed for the independent test–set evaluation.

Figures 5.6 and 5.7 show the validation *TSS* as a function of the decision threshold for the selected static– and dynamic–undersampling models, respectively. In both cases, the *TSS* curves exhibit broad near–optimal regions rather than isolated sharp peaks, indicating that the validation skill is relatively stable over a range of threshold values. Therefore, the selected thresholds should be interpreted as empirical operating points within near–optimal regions, rather than as points at which the model suddenly changes behavior or as uniquely meaningful boundary values.

Quantitative results on threshold stability are reported in Table 5.4, which lists the threshold intervals over which the validation *TSS* remains within 0.01 of its maximum value. The static *24H24H36M* and *6H6H12M* models show the broadest near–optimal intervals, with widths of 0.247 and 0.251, respectively, indicating the greatest robustness to the exact threshold choice. In contrast, the *12H12H12M* models have much narrower intervals for both static and dynamic undersampling, with widths of 0.056 and 0.060, respectively. This indicates that, although *12H12H12M* models achieve the highest validation *TSS* their performance is more sensitive to the exact threshold value.

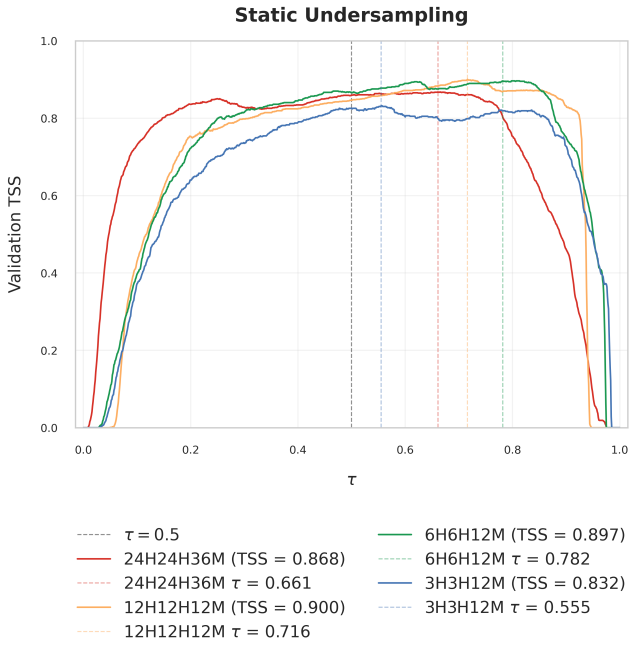


FIGURE 5.6: Validation *TSS* as a function of the decision threshold for the selected static–undersampling models. The solid curves show the validation *TSS* obtained across threshold values, the colored vertical dashed lines indicate the *TSS*–optimal thresholds selected on the validation set, and the gray vertical dashed line marks the default threshold ($\tau = 0.5$).

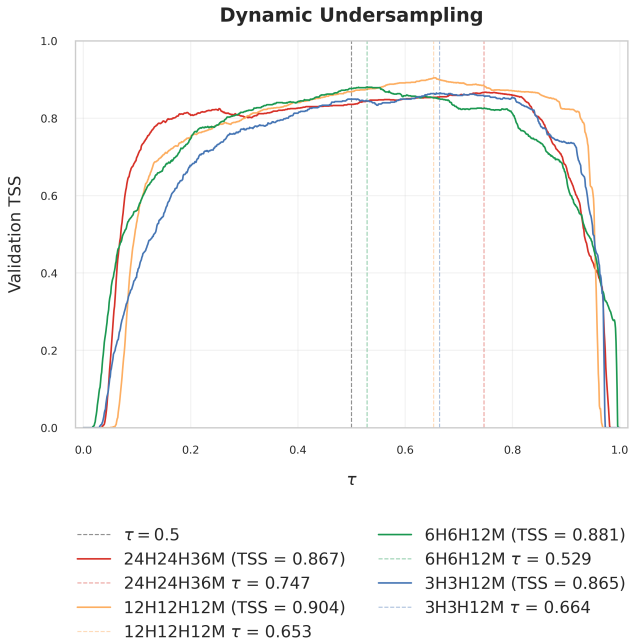


FIGURE 5.7: Validation *TSS* as a function of the decision threshold for the selected dynamic–undersampling models. The solid curves show the validation *TSS* obtained across threshold values, the colored vertical dashed lines indicate the *TSS*–optimal thresholds selected on the validation set, and the gray vertical dashed line marks the default threshold ($\tau = 0.5$).

For static undersampling, the *TSS*–optimal thresholds were above the default value ($\tau = 0.5$), ranging from 0.555 for the *3H3H12M* model to 0.782 for the *6H6H12M* model (see Table 5.4 and Table 5.5). The highest validation *TSS* was obtained by the *12H12H12M* model, with $TSS = 0.900$ at a threshold of 0.716. This model also had a balanced *Recall* and *Specificity* of 0.951 and 0.949, respectively. The *6H6H12M* model reached a comparable *TSS* of 0.897, but with a different balance between *Recall* and *Specificity*, corresponding to 0.920 for *Recall* and 0.977 for *Specificity*.

The dynamic undersampling selected thresholds that were also above the default value of ($\tau = 0.5$), ranging from 0.529 for the *6H6H12M* model to 0.747 for the *24H24H36M* model (see Table 5.4 and Table 5.5). In this case also, the *12H12H12M* model was best according to validation *TSS*, which reached $TSS = 0.904$ at a threshold of 0.653 and had a balanced *Recall* and *Specificity* of 0.951 and 0.953, respectively. The remaining metrics are reported in Table 5.5. As in the preceding chapter, these values are reported only for reference, since the final results require evaluation on the independent test set.

The effect of threshold optimization relative to the default threshold ($\tau = 0.5$) is reported in Table 5.6. Threshold optimization increased the validation *TSS* for all temporal configurations and for both undersampling strategies. For static undersampling, the largest improvement was obtained by *12H12H12M* model, with $\Delta TSS = 0.054$. The *6H6H12M* model also showed a clear improvement, with $\Delta TSS = 0.030$, whereas the gains for *24H24H36M* and *3H3H12M* were more modest. For dynamic undersampling, the largest improvements were obtained by the *12H12H12M* and *24H24H36M* models, with $\Delta TSS = 0.036$ and $\Delta TSS = 0.031$, respectively. The *3H3H12M* model showed a smaller improvement, with $\Delta TSS = 0.016$, while the improvement for the *6H6H12M* model was the smallest, with $\Delta TSS = 0.004$, indicating that the default threshold was already close to the validation–optimal operating point for this model.

Because all optimized thresholds were higher than the default value ($\tau = 0.5$), threshold optimization generally increased *Specificity* while slightly decreasing, or leaving unchanged, *Recall*. This indicates that the optimized operating points tended to reduce the *FPR* rather than increase the *TPR*. This trade–off is consistent with the use of *TSS* as the optimization criterion, since *TSS* evaluates *Sensitivity* and *Specificity* jointly. This effect is particularly clear for the static *6H6H12M* model, where *Specificity* increased from 0.908 to 0.977, while *Recall* decreased from 0.959 to 0.920.

Threshold optimization produced consistent, although mostly modest, improvements in validation *TSS*. The main effect was to move the operating point toward higher thresholds. The optimized thresholds obtained in this subsection define the operating points used for the final test–set evaluation.

Undersampling Strategy	Temporal	TSS–Optimal Threshold	Maximum TSS	Threshold Interval	Interval Width
Static	24H24H36M	0.661	0.868	[0.481, 0.728]	0.247
	12H12H12M	0.716	0.900	[0.684, 0.740]	0.056
	6H6H12M	0.782	0.897	[0.591, 0.842]	0.251
	3H3H12M	0.555	0.832	[0.470, 0.579]	0.109
Dynamic	24H24H36M	0.747	0.867	[0.681, 0.807]	0.126
	12H12H12M	0.653	0.904	[0.620, 0.680]	0.060
	6H6H12M	0.529	0.881	[0.480, 0.561]	0.081
	3H3H12M	0.664	0.865	[0.625, 0.758]	0.133

TABLE 5.4: Threshold sensitivity for the selected models. The threshold interval indicates the range over which the validation *TSS* remains within 0.01 of its maximum value.

Undersampling Strategy	Temporal Configuration	Threshold	TSS	Accuracy	Precision	Recall	Specificity	F1–score	FPR	FNR
Static	24H24H36M	0.661	0.868	0.934	0.975	0.891	0.977	0.931	0.023	0.109
	12H12H12M	0.716	0.900	0.950	0.949	0.951	0.949	0.950	0.051	0.049
	6H6H12M	0.782	0.897	0.948	0.975	0.920	0.977	0.947	0.023	0.080
	3H3H12M	0.555	0.832	0.916	0.892	0.946	0.886	0.918	0.114	0.054
Dynamic	24H24H36M	0.747	0.867	0.934	0.963	0.902	0.966	0.931	0.034	0.098
	12H12H12M	0.653	0.904	0.952	0.953	0.951	0.953	0.952	0.047	0.049
	6H6H12M	0.529	0.881	0.941	0.930	0.953	0.928	0.941	0.072	0.047
	3H3H12M	0.664	0.865	0.933	0.950	0.913	0.952	0.931	0.048	0.087

TABLE 5.5: Validation–set performance of the selected models at the *TSS*–optimal threshold. The optimal threshold is selected by maximizing validation *TSS*; when multiple thresholds achieve the same maximum *TSS*, the lowest value among the tied optima is selected. All remaining metrics are computed at the selected threshold.

Undersampling Strategy	Temporal Configuration	TSS			Recall			Specificity		
		$\tau_{0.5}$	τ_{opt}	Δ	$\tau_{0.5}$	τ_{opt}	Δ	$\tau_{0.5}$	τ_{opt}	Δ
Static	24H24H36M	0.859	0.868	0.009	0.909	0.891	-0.018	0.950	0.977	0.027
	12H12H12M	0.846	0.900	0.054	0.951	0.951	0.000	0.895	0.949	0.054
	6H6H12M	0.867	0.897	0.030	0.959	0.920	-0.040	0.908	0.977	0.069
	3H3H12M	0.826	0.832	0.006	0.970	0.946	-0.024	0.856	0.886	0.030
Dynamic	24H24H36M	0.836	0.867	0.031	0.912	0.902	-0.011	0.923	0.966	0.042
	12H12H12M	0.868	0.904	0.036	0.951	0.951	0.000	0.917	0.953	0.036
	6H6H12M	0.877	0.881	0.004	0.955	0.953	-0.002	0.922	0.928	0.007
	3H3H12M	0.849	0.865	0.016	0.941	0.913	-0.029	0.908	0.952	0.044

TABLE 5.6: Effect of threshold optimization on validation–set performance for the selected models. The default threshold is $\tau = 0.5$, while the optimized threshold is selected by maximizing validation *TSS*. Differences are computed as optimized–threshold performance minus default–threshold performance.

5.4.3 Test–Set Forecasting Performance

Once the best model had been selected for each sequence configuration and undersampling strategy based on the minimum validation loss, and the decision threshold had been optimized according to the maximum validation–set *TSS*, the selected models were evaluated on the independent test set. No additional tuning was performed using test–set information. Table 5.7 reports the confusion–matrix values and classification metrics obtained for each sequence configuration and undersampling strategy using the corresponding validation–optimized threshold.

Across all selected models, the best test–set performance was obtained by the static *12H12H12M* model, which reached $TSS = 0.892$, with a *Recall* of 0.950 and a *Specificity* of 0.942. This model provided the most balanced *Recall–Specificity* performance among the evaluated configurations. The dynamic *12H12H12M* model achieved the second–highest *TSS*, 0.856, also with a relatively balanced *Recall* and *Specificity* of 0.916 and 0.940, respectively. These results indicate that the *12H12H12M* sequence configuration provided the strongest overall forecasting performance on the independent test set for both undersampling strategies.

The remaining models show different trade–offs between *Recall* and *Specificity*. For static undersampling, the *24H24H36M* model achieved high *Specificity*, 0.968, but lower *Recall*, 0.854, resulting in $TSS = 0.822$. The *3H3H12M* model achieved higher *Recall*, 0.924, but lower *Specificity*, 0.889, resulting in $TSS = 0.814$. The *6H6H12M* model showed a different pattern, with high *Specificity*, 0.954, but lower *Recall*, 0.839, giving $TSS = 0.793$. Thus, within the static–undersampling models, the shorter *3H3H12M* temporal configuration tended to favor the correct detection of flare–productive sequences, whereas the longer and intermediate temporal configurations favored false–alarm reduction.

For the dynamic undersampling, the best performance after the *12H12H12M* model was obtained by the *3H3H12M* model, with $TSS = 0.839$, $Recall = 0.901$, and $Specificity = 0.938$. The dynamic *6H6H12M* model reached $TSS = 0.810$, with $Recall = 0.891$ and $Specificity = 0.919$, while the lowest dynamic *TSS*, 0.792, was obtained by the dynamic *24H24H36M* model, mainly because of its lower *Recall*, 0.842, despite maintaining high *Specificity*, 0.950. Overall, the dynamic–undersampling models showed a more balanced relation between *Recall* and *Specificity*.

A direct comparison between static and dynamic undersampling based on *TSS* is reported in Table 5.8. Static undersampling performed better for the two longer sequence configurations, with dynamic–minus–static differences of $\Delta TSS = -0.030$ for *24H24H36M* and $\Delta TSS = -0.036$ for *12H12H12M*. In contrast, dynamic undersampling improved performance for the shorter sequence configurations, with $\Delta TSS = 0.017$ for *6H6H12M* and $\Delta TSS = 0.025$ for *3H3H12M*. In all cases, however, the differences were modest.

			Confusion Matrix				Metrics							
Undersampling Strategy	**Temporal Configuration**	**Threshold**	**TP**	**TN**	**FP**	**FN**	**TSS**	**Acc.**	**Prec.**	**Rec.**	**Spec.**	**F1**	**FPR**	**FNR**
Static	24H24H36M	0.661	1546	1752	58	264	0.822	0.911	0.964	0.854	0.968	0.906	0.032	0.146
	12H12H12M	0.716	1485	1472	91	78	0.892	0.946	0.942	0.950	0.942	0.946	0.058	0.050
	6H6H12M	0.782	1034	1175	57	198	0.793	0.897	0.948	0.839	0.954	0.890	0.046	0.161
	3H3H12M	0.555	734	706	88	60	0.814	0.907	0.893	0.924	0.889	0.908	0.111	0.076
Dynamic	24H24H36M	0.747	1524	1720	90	286	0.792	0.896	0.944	0.842	0.950	0.890	0.050	0.158
	12H12H12M	0.653	1432	1469	94	131	0.856	0.928	0.938	0.916	0.940	0.927	0.060	0.084
	6H6H12M	0.529	1098	1132	100	134	0.810	0.905	0.917	0.891	0.919	0.904	0.081	0.109
	3H3H12M	0.664	715	745	49	79	0.839	0.919	0.936	0.901	0.938	0.918	0.062	0.099

TABLE 5.7: Confusion–matrix counts and classification metrics computed on the test set for the selected models using thresholds optimized on the validation set.

	TSS			Recall			Specificity		
Temporal Configuration	**Static**	**Dynamic**	**Δ**	**Static**	**Dynamic**	**Δ**	**Static**	**Dynamic**	**Δ**
24H24H36M	0.822	0.792	-0.030	0.854	0.842	-0.012	0.968	0.950	-0.018
12H12H12M	0.892	0.856	-0.036	0.950	0.916	-0.034	0.942	0.940	-0.002
6H6H12M	0.793	0.810	0.017	0.839	0.891	0.052	0.954	0.919	-0.035
3H3H12M	0.814	0.839	0.025	0.924	0.901	-0.024	0.889	0.938	0.049

TABLE 5.8: Direct comparison between static and dynamic undersampling on the test set. For each temporal configuration, metric differences are computed as dynamic minus static for *TSS*, *Recall*, and *Specificity*.

The underlying *Recall–Specificity* trade–off for the improvement of dynamic over static undersampling varied across the shorter sequence configurations; for the *6H6H12M* sequence configuration, dynamic undersampling increased *Recall* but reduced *Specificity*, whereas for the *3H3H12M* sequence configuration, it improved *Specificity* while slightly reducing *Recall*.

The ranking of the selected models according to test–set *TSS* is reported in Table 5.9. The two highest–ranked configurations were the static *12H12H12M* model, followed by the dynamic *12H12H12M* model. This indicates that, among the evaluated sequence configurations, *12H12H12M* provided the strongest test–set performance for both undersampling strategies. However, because each sequence configuration combines a specific observational window, forecasting window, and cadence, this result should not be interpreted as an isolated effect of the observational window or forecasting window, nor as evidence for a physically preferred precursor timescale or forecasting horizon. The ranking also shows that dynamic–undersampling models remained competitive and ranked above the corresponding static–undersampling model for the shorter *6H6H12M* and *3H3H12M* temporal configurations. The differences between static and dynamic undersampling were modest, and the results do not indicate a clear universal advantage of either strategy. Therefore, the preferred undersampling strategy appears to be configuration–dependent rather than universally fixed.

Rank	Undersampling Strategy	Temporal Configuration	τ	TSS	Rec.	Spec.
1	Static	12H12H12M	0.716	0.892	0.950	0.942
2	Dynamic	12H12H12M	0.653	0.856	0.916	0.940
3	Dynamic	3H3H12M	0.664	0.839	0.901	0.938
4	Static	24H24H36M	0.661	0.822	0.854	0.968
5	Static	3H3H12M	0.555	0.814	0.924	0.889
6	Dynamic	6H6H12M	0.529	0.810	0.891	0.919
7	Static	6H6H12M	0.782	0.793	0.839	0.954
8	Dynamic	24H24H36M	0.747	0.792	0.842	0.950

TABLE 5.9: Ranking of the selected models according to test–set *TSS*. *Recall* and *Specificity* are reported to show the *Recall–Specificity* trade–off obtained at each validation–optimized threshold.

Taken together, these results identify the static *12H12H12M* model as the best–performing classifier according to test–set *TSS*, while showing that the relative benefit of static and dynamic undersampling depends on the sequence configuration.

5.4.4 Discrimination and Probability Separation

In the previous subsection, we evaluated the performance of the selected models on the independent test set after applying the validation–optimized decision thresholds. In this subsection, we complement that analysis by examining two related properties of the model outputs. First, we assess threshold–independent discrimination, that is, whether the models can rank flare–productive sequences above non–flaring sequences independently of a particular operating threshold. To this end, we visually inspect the ROC curves and report *ROC–AUC* and *PR–AUC*, which summarize performance over the full range of possible decision thresholds. Second, we inspect the predicted–probability distributions to evaluate whether positive and negative sequences occupy distinct regions of probability space. We also report summary statistics of probability separation based on the class–wise mean and median predicted probabilities.

Figures 5.8 and 5.9 show the test–set ROC curves for the selected static and dynamic–undersampling models, respectively. All ROC curves lie well above the no–skill diagonal, indicating strong discrimination between positive and negative sequences. The markers indicate the operating points corresponding to the validation–optimized thresholds used in the previous subsection. These operating points are located in regions with high *TPR* and low *FPR*, confirming that the optimized thresholds select favorable regions of the ROC curves.

The corresponding threshold–independent discrimination metrics are reported in Table 5.10. Across all selected models, *ROC–AUC* values ranged from 0.962 to 0.981, corresponding to a relative spread of approximately 2.0%, while *PR–AUC* values ranged from 0.968 to 0.980, corresponding to a relative spread of approximately 1.2%. The dynamic *12H12H12M* model obtained the highest values for both *ROC–AUC* and *PR–AUC*. These results show that all finalized models retained strong threshold–independent discrimination on the independent test set.

Although the static *12H12H12M* model achieved the best threshold–based forecasting performance in the previous subsection, its *ROC–AUC* was 0.973, equal to that of the static *24H24H36M* model. This value was slightly higher than the lowest static *ROC–AUC*, obtained by the *6H6H12M* model, and only slightly lower than the highest static *ROC–AUC*, obtained by the *3H3H12M* model. However, because these values were very similar across static sequence configurations, ranging from 0.971 to 0.976 for, this ranking should not be overinterpreted. A similar reasoning applies to *PR–AUC*, whose values also varied only marginally across the static models.

For the dynamic–undersampling models, discrimination was also strong, with the values showing a wider range of variation than for the static–undersampling models. The dynamic *12H12H12M* model, which achieved the best threshold–based performance among the dynamic–undersampling models in the previous subsection, also obtained the highest dynamic *ROC–AUC*, 0.981, and tied for the highest dynamic *PR–AUC*, 0.980, with the dynamic *3H3H12M* model. The dynamic *3H3H12M* model also ranked second among

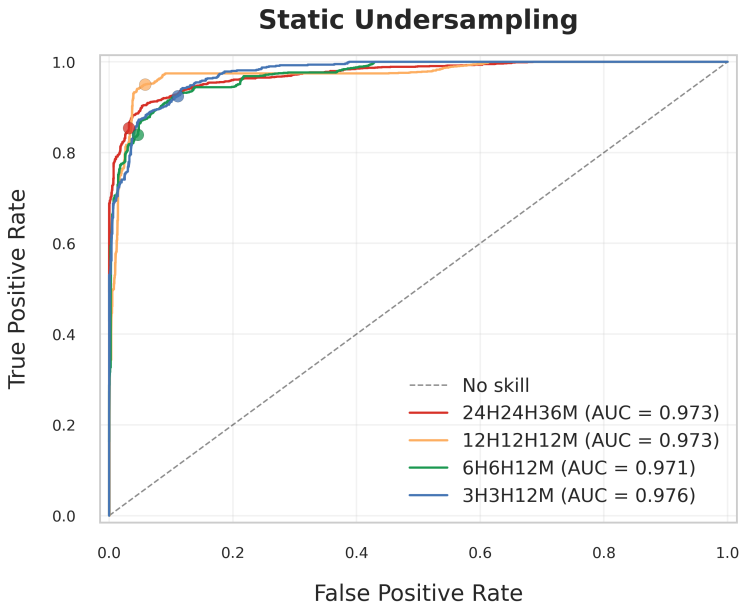


FIGURE 5.8: Test–set ROC curves for the selected static–undersampling models. The diagonal dashed line indicates the no–skill reference, and the markers indicate the operating points obtained using validation–optimized thresholds.

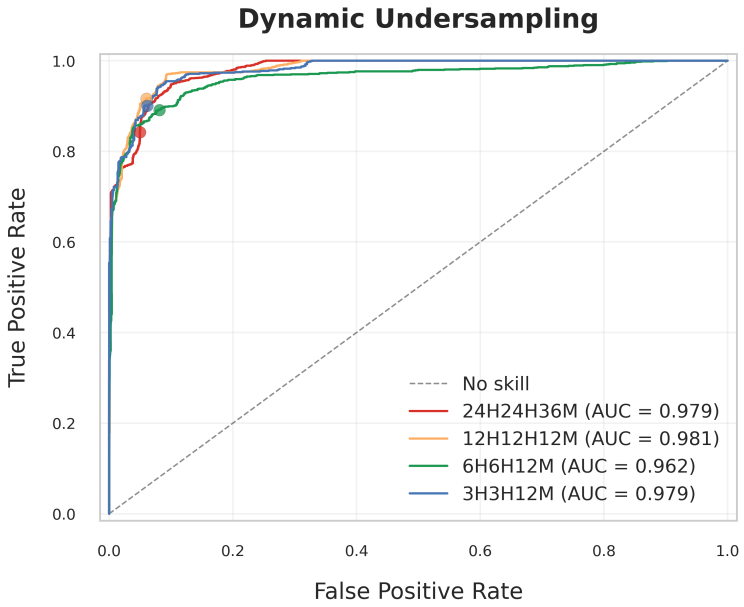


FIGURE 5.9: Test–set ROC curves for the selected dynamic–undersampling models. The diagonal dashed line indicates the no–skill reference, and the markers indicate the operating points obtained using validation–optimized thresholds.

the dynamic models according to test–set *TSS*, indicating consistency between its threshold–based performance and its threshold–independent discrimination.

The test–set predicted–probability distributions for positive and negative sequences are shown in Figure 5.10. The distributions are colored according to the original test–set labels, with positive sequences corresponding to $y_i = 1$ and negative sequences corresponding to $y_i = 0$. The vertical dashed lines indicate the validation–optimized thresholds used to convert predicted probabilities into binary forecasts, but these thresholds are shown only as reference operating points.

The probability–separation statistics reported in Table 5.10 provide a numerical counterpart of the information shown in Figure 5.10. They are computed from the same true–label grouping. For each model, the predicted probabilities $\hat{p}_i$ are divided into the set of probabilities assigned to truly positive sequences,

$$P_+ = \{\hat{p}_i : y_i = 1\}, \tag{5.21}$$

and the set of probabilities assigned to truly negative sequences,

$$P_- = \{\hat{p}_i : y_i = 0\}. \tag{5.22}$$

The class–wise mean probabilities are then computed as

$$\text{Mean}_+ = \frac{1}{N_+} \sum_{i:y_i=1} \hat{p}_i, \qquad \text{Mean}_- = \frac{1}{N_-} \sum_{i:y_i=0} \hat{p}_i, \tag{5.23}$$

and their difference is reported as

$$\Delta\text{Mean} = \text{Mean}_+ - \text{Mean}_-. \tag{5.24}$$

Similarly, Median_+ and Median_- are computed as the medians of P_+ and P_-, respectively, and the median separation is defined as

$$\Delta\text{Median} = \text{Median}_+ - \text{Median}_-. \tag{5.25}$$

Thus, the summary statistics quantify how far apart the predicted probabilities of the true positive and true negative sequences are in probability space, independently of the threshold used for binary classification.

The probability histograms show that, in all configurations, the negative sequences are concentrated near low predicted probabilities, while the positive sequences are concentrated near high predicted probabilities. They also show that some overlap remains between the two classes, especially in intermediate probability regions. This overlap corresponds to sequences for which the model assigns less certain probabilities and explains why the final forecasting metrics remain sensitive to the decision threshold. Nevertheless, the strong concentration of positive sequences near high probabilities and negative sequences near low probabilities confirms that the selected models learned a meaningful probabilistic separation between flare–productive and non–flaring sequences.

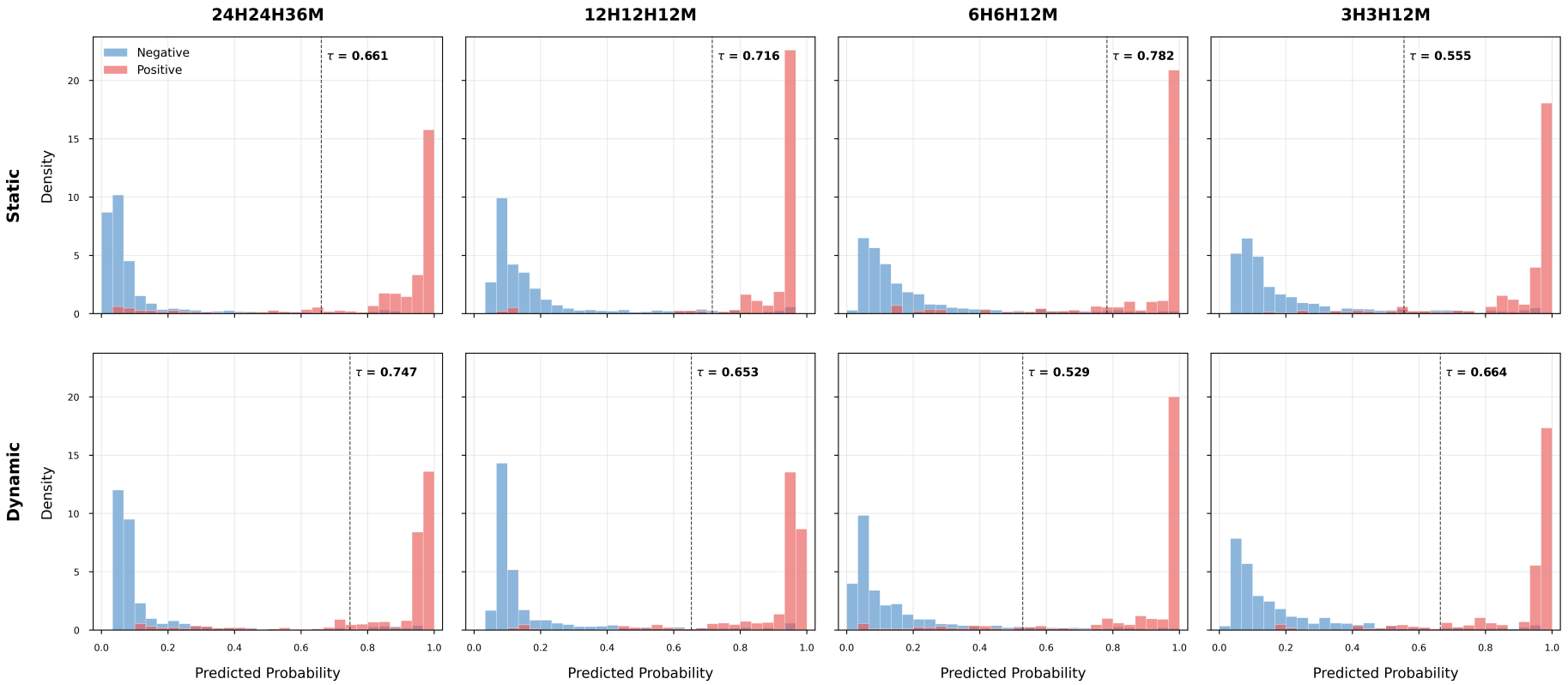


FIGURE 5.10: Test–set predicted–probability distributions for the selected static– and dynamic–undersampling models. Columns correspond to the four temporal configurations, while rows correspond to the undersampling strategy. Each panel shows the probability distributions for negative and positive sequences, with the vertical dashed line indicating the validation–optimized threshold.

		Discrimination		Mean Probability			Median Probability			Samples	
Undersampling Strategy	**Temporal Configuration**	**ROC–AUC**	**PR–AUC**	**Mean$_+$**	**Mean$_-$**	**ΔMean**	**Median$_+$**	**Median$_-$**	**ΔMedian**	N_+	N_-
Static	24H24H36M	0.973	0.978	0.856	0.109	0.747	0.971	0.050	0.921	1810	1810
	12H12H12M	0.973	0.971	0.899	0.205	0.695	0.941	0.119	0.822	1563	1563
	6H6H12M	0.971	0.972	0.884	0.204	0.680	0.975	0.118	0.857	1232	1232
	3H3H12M	0.976	0.975	0.901	0.218	0.683	0.975	0.121	0.854	794	794
Dynamic	24H24H36M	0.979	0.978	0.868	0.144	0.724	0.964	0.074	0.890	1810	1810
	12H12H12M	0.981	0.980	0.893	0.179	0.713	0.962	0.097	0.865	1563	1563
	6H6H12M	0.962	0.968	0.874	0.165	0.709	0.988	0.073	0.915	1232	1232
	3H3H12M	0.979	0.980	0.895	0.196	0.698	0.971	0.115	0.856	794	794

TABLE 5.10: Threshold–independent discrimination and probability–separation metrics computed on the test set for the selected models. *ROC–AUC* and *PR–AUC* quantify discrimination, while the mean and median probability differences summarize the separation between truly positive and truly negative sequences. The sample columns report the number of positive and negative sequences for each temporal configuration for reference.

Across all models, the mean–probability separation ranged from 0.680 to 0.747, corresponding to a relative spread of approximately 9.9%, and the median–probability separation ranged from 0.822 to 0.921, corresponding to a relative spread of approximately 12.0%. The largest mean–probability separation was obtained by the static *24H24H36M* model, with $\Delta\text{Mean} = 0.747$, and the largest median–probability separation was obtained by the same model, with $\Delta\text{Median} = 0.921$. The smallest mean–probability separation was obtained by the static *6H6H12M* model, with $\Delta\text{Mean} = 0.680$, whereas the smallest median–probability separation was obtained by the static *12H12H12M* model, with $\Delta\text{Median} = 0.822$. These results indicate that the positive and negative probability distributions are generally well separated. The difference between the mean– and median–based separations reflects the shape of the predicted–probability distributions, since the mean is affected by outlying probabilities whereas the median provides a more robust summary of the typical separation between the two classes.

The analysis conducted in the previous subsection addresses the question of how well the finalized models perform after selecting a specific decision threshold. In contrast, this subsection addresses two related but distinct questions: first, whether the models can rank positive sequences above negative sequences independently of the chosen operating threshold; and second, whether the predicted probabilities of positive and negative sequences occupy distinct regions of probability space. The results show that all finalized models provide strong discrimination, while the probability histograms and the mean– and median–based separation statistics show that positive and negative sequences are generally well separated in probability space, although some overlap remains in intermediate–probability regions. The best–performing configuration depends on the metric under consideration, confirming that *ROC–AUC*, *PR–AUC*, and summary statistics of probability–distribution separation provide complementary views of model behavior. In particular, the static *12H12H12M* model was the best threshold–based classifier according to test–set *TSS*, whereas the dynamic *12H12H12M* model showed the strongest threshold–independent discrimination.

5.4.5 Calibration Analysis

The previous subsection showed that the selected models provide strong discrimination between flare–productive and non–flaring sequences. However, strong discrimination does not necessarily imply that the predicted probabilities are well calibrated. Calibration evaluates whether predicted probabilities correspond to observed event frequencies. For example, among samples assigned a predicted probability close to 0.7, a well–calibrated model should produce an observed positive fraction close to 0.7. If sequences assigned a predicted probability close to 0.8 (0.6) produce flares in about 0.6 (0.8) of the cases, the model is overconfident (underconfident), even if it remains discriminative. This subsection therefore evaluates the reliability of the predicted probabilities on the independent test set, that is, whether these values can be interpreted as reliable estimates of flare occurrence.

Because the test set is balanced by construction, the test–set results correspond to a 50% flare prevalence rather than to the much lower prevalence expected in an operational forecasting setting. *TSS* and *ROC–AUC* are unaffected by class prevalence, whereas *Accuracy*, *Precision*, *F1–score*, *PR–AUC*, *Brier score*, *ECE*, and *MCE* depend on the underlying class distribution. Therefore, the reliability diagrams and calibration metrics should be interpreted as comparisons among the models under this controlled evaluation setting, rather than as estimates of probability calibration under operational flare occurrence rates.

Figures 5.11 and 5.12 show the test–set reliability diagrams for the static– and dynamic–undersampling models, respectively. The predicted probabilities were divided into 10 equally spaced bins over the interval $[0, 1]$. The dashed diagonal line represents perfect calibration. Points above the diagonal indicate probability bins in which the observed positive fraction is higher than the mean predicted probability, corresponding to underconfidence. Points below the diagonal indicate bins in which the observed positive fraction is lower than the mean predicted probability, corresponding to overconfidence.

For the static–undersampling models, the reliability curves generally follow the diagonal trend, indicating broadly reasonable calibration. This behavior is consistent with the *ECE* values reported in the legends, which remain below 0.1, ranging from 0.0212 to 0.0805 for the *24H24H36M* and *12H12H12M* sequence configurations, respectively. According to *ECE*, the static–undersampling models are ranked, from best to worst calibrated, as *24H24H36M*, *6H6H12M*, *3H3H12M*, and *12H12H12M*. The deviations from perfect calibration are mainly visible in intermediate probability bins, where some models lie above or below the diagonal. This pattern also reflects the fact that most predicted probabilities are tightly clustered near 0 and 1, while the intermediate bins contain fewer samples and are therefore more sensitive to local deviations.

For the dynamic undersampling models, the reliability curves also remain close to the perfect–calibration diagonal, although deviations are again visible in some intermediate probability bins. Compared with the static undersampling models, the *ECE* values are less spread out, ranging from 0.0225 to 0.0574 for the *6H6H12M* and *24H24H36M* sequence configurations, respectively, indicating more uniform average calibration errors. In this case, the best–calibrated model according to *ECE* is *6H6H12M*, followed by *12H12H12M*, *3H3H12M*, and *24H24H36M*.

The reliability diagrams also show that calibration deviations are not systematically in one direction across all models and probability bins. For example, the static *24H24H36M* model, which has the lowest *ECE* among the static models, fluctuates around the diagonal and remains comparatively well calibrated over much of the probability range, whereas the dynamic *24H24H36M* model lies above the diagonal in several intermediate bins, indicating underconfidence in those regions. The dynamic *6H6H12M* model, which has the lowest *ECE* among the dynamic models, remains close to the diagonal across much of the probability range. In contrast, the static *6H6H12M* model shows

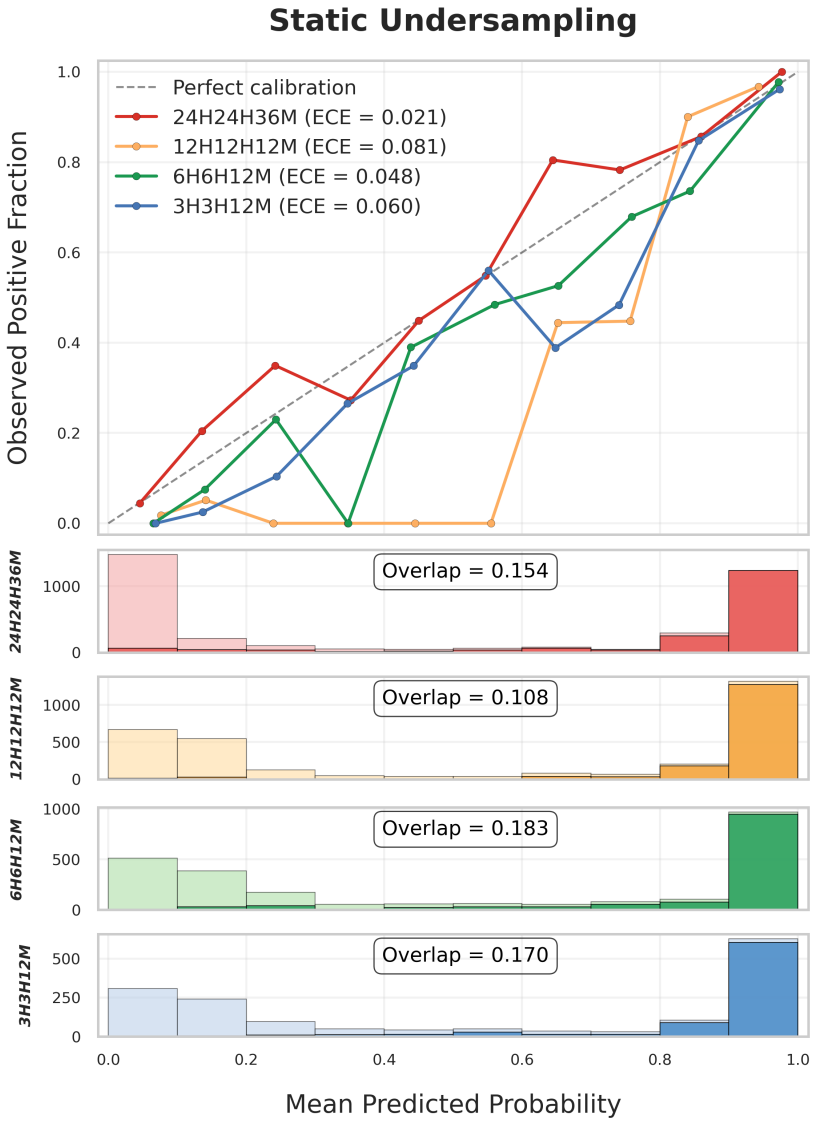


FIGURE 5.11: Test–set reliability diagram for the selected static–undersampling models. The dashed diagonal line indicates perfect calibration. Each point represents a non–empty probability bin, with the x–axis corresponding to the mean predicted probability and the y–axis corresponding to the observed positive fraction. The histograms show the distribution of predicted probabilities for each temporal configuration.

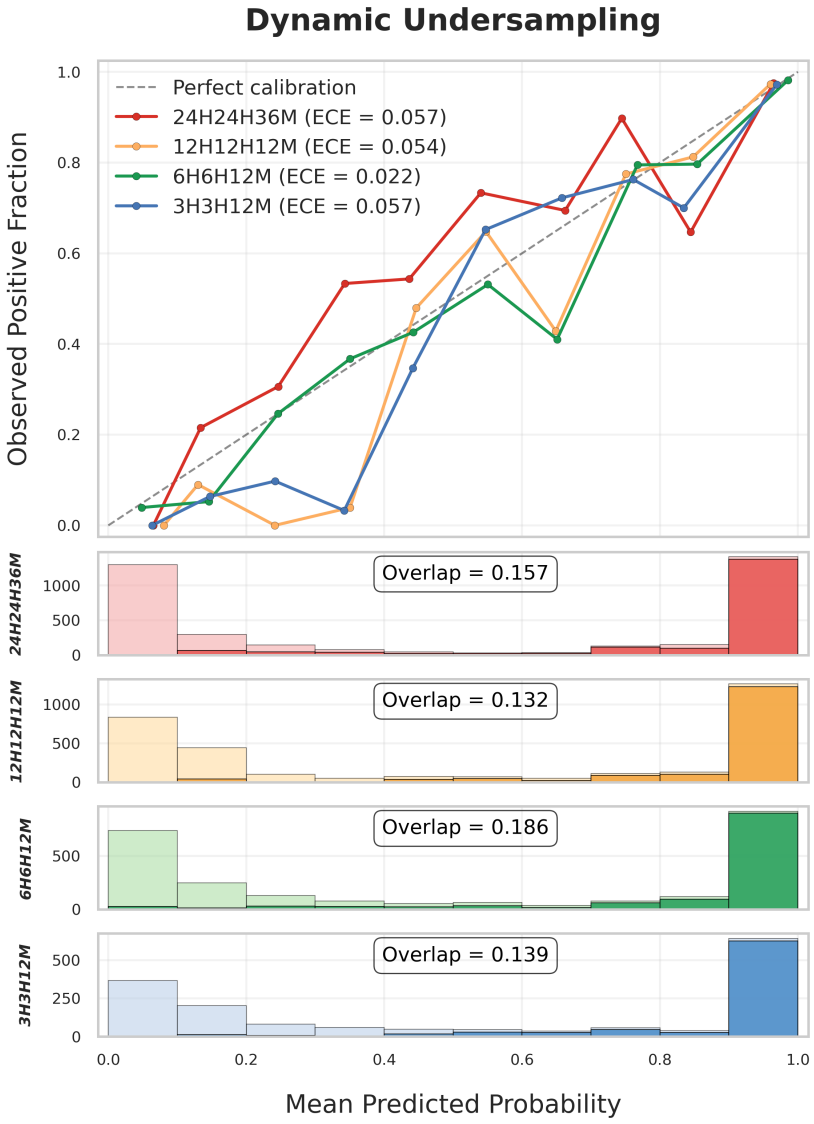


FIGURE 5.12: Test–set reliability diagram for the selected dynamic–undersampling models. The dashed diagonal line indicates perfect calibration. Each point represents a non–empty probability bin, with the x–axis corresponding to the mean predicted probability and the y–axis corresponding to the observed positive fraction. The histograms show the distribution of predicted probabilities for each temporal configuration.

more visible deviations, with several probability bins lying below the diagonal, indicating overconfidence.

The *12H12H12M* and *3H3H12M* models show mixed behavior: in the static undersampling case, the reliability curves tend to remain below or close to the diagonal in several probability bins, indicating a tendency toward overconfidence, while in the dynamic undersampling case, some fluctuations are observed around the diagonal, with no single systematic direction across the full probability range.

The histograms below the reliability diagrams provide a complementary view by showing how the predicted probabilities are distributed across bins. They show that most samples are concentrated near the extremes of the probability range, with many negative sequences assigned low probabilities and many positive sequences assigned high probabilities. Consequently, intermediate probability bins often contain fewer samples, making the corresponding observed positive fractions more sensitive to small–number fluctuations.

The plots also report the *Overlap* between the positive and negative predicted–probability distributions, computed as defined in Equation 5.15. The effect of sparsely populated intermediate bins is particularly visible for the static *12H12H12M* model. As shown in Figure 5.11, some intermediate bins contain few or no positive sequences, producing calibration points that depart more strongly from the diagonal. This model is the worst–calibrated according to *ECE*, but it also has the lowest distribution overlap, with an *Overlap* of 0.108, indicating strong separation between the positive and negative predicted–probability distributions. Thus, the histogram information supports the interpretation that the largest calibration deviations occur mainly in sparsely populated probability regions, rather than in the high–density regions near 0 and 1.

Table 5.11 reports the *Brier score*, *ECE*, and *MCE* for all selected models, as defined in Section 5.3.2. Lower values indicate better calibration. The *Brier score* evaluates the overall probabilistic accuracy of the predictions, while *ECE* and *MCE* summarize the average and maximum bin–wise calibration gaps, respectively. Thus, *ECE* reflects the typical calibration error across probability bins, whereas *MCE* highlights the largest localized deviation from perfect calibration.

Across all models, *Brier score* ranged from 0.0558 to 0.0705 for the dynamic *12H12H12M* model and the static *6H6H12M* model, respectively. The differences in *Brier score* were small across models, indicating broadly similar probabilistic accuracy according to this metric. The *ECE* and *MCE* values support the visual interpretation provided by the reliability diagrams. Table 5.11 also shows that probabilistic accuracy, average calibration error, and localized calibration error do not always rank the models in the same way. For example, the static *12H12H12M* model has the second–lowest *Brier score*, but the highest *MCE*, indicating that its overall probabilistic accuracy is good while one probability region shows a larger localized calibration deviation.

The calibration analysis shows that the selected models produce probability estimates that are broadly meaningful, but not uniformly calibrated

Undersampling Strategy	Temporal Configuration	Brier Score	ECE	MCE
Static	24H24H36M	0.0608	0.0212	0.1604
	12H12H12M	0.0594	0.0805	0.5548
	6H6H12M	0.0705	0.0481	0.3476
	3H3H12M	0.0687	0.0602	0.2594
Dynamic	24H24H36M	0.0620	0.0574	0.1978
	12H12H12M	0.0558	0.0541	0.3113
	6H6H12M	0.0701	0.0225	0.2406
	3H3H12M	0.0612	0.0568	0.3093

TABLE 5.11: Test–set calibration metrics for the selected models. The *Brier score* measures the mean squared error of the predicted probabilities, while *ECE* and *MCE* quantify the expected and maximum calibration gaps across probability bins, respectively. Lower values indicate better calibration.

across all probability regions. The reliability diagrams and *ECE* values indicate generally small average calibration errors, while the histograms show that the most stable probability regions are those near 0 and 1, where most predictions are concentrated. Larger deviations appear mainly in intermediate probability bins, where fewer samples are available and the observed positive fraction is more sensitive to local fluctuations.

The comparison between undersampling strategies does not identify a uniformly better calibrated approach. Static undersampling achieved lower *Brier score* for some sequence configurations, suggesting slightly better overall probabilistic accuracy in those cases. In contrast, dynamic undersampling produced a narrower range of *ECE* values and included the model with the lowest *MCE*, indicating more uniform average calibration errors and smaller localized deviations for some configurations. Therefore, the apparent calibration advantage depends on the metric considered: static undersampling is favored by the *Brier score* in some cases, whereas dynamic undersampling appears more stable in terms of bin–wise calibration errors.

In conclusion, calibration provides information that is complementary to both threshold–based classification performance and threshold–independent discrimination: the best classifier according to test–set *TSS* is not necessarily the model with the most reliable probability estimates across all probability ranges.

5.4.6 Summary of Findings

A complete evaluation of the proposed sequence–based flare–forecasting framework is presented in this section. The first part of the analysis consisted

of the model–selection stage, in which one model configuration was identified for each combination of temporal configuration and undersampling strategy. The hyperparameter search explored four LSTM hidden sizes, 32, 64, 128, and 256, and four learning rates, 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. The selected model in each case was the one that achieved the minimum validation loss. The decision thresholds were then optimized on the validation set by maximizing *TSS* and kept fixed for the independent test–set evaluation. This procedure selected eight models, corresponding to the four temporal sequence configurations and the two undersampling strategies. The remainder of the analysis was guided by the questions formulated at the beginning of the section; the following summary revisits those questions and synthesizes the main findings.

1. Which temporal configuration provides the strongest forecasting performance for major flare prediction?

The best test–set forecasting performance was obtained by the *12H12H12M* temporal configuration. In particular, the static *12H12H12M* model achieved the highest test–set *TSS*, with $TSS = 0.892$, together with a balanced *Recall* of 0.950 and *Specificity* of 0.942. The dynamic *12H12H12M* model ranked second, with $TSS = 0.856$, *Recall* of 0.916, and *Specificity* of 0.940. Thus, among the evaluated temporal configurations, *12H12H12M* provided the strongest forecasting performance for both undersampling strategies. However, because each sequence configuration combines a different observational window, forecasting window, and temporal cadence, this result should be interpreted as identifying the best–performing evaluated configuration, rather than as evidence for a physically preferred precursor timescale or forecasting horizon.

2. How does dynamic undersampling compare with static undersampling across the evaluated temporal configurations?

The comparison between static and dynamic undersampling did not show a universal advantage of either strategy. Static undersampling produced the strongest overall classifier, namely the static *12H12H12M* model, and performed better for the two longer temporal configurations, *24H24H36M* and *12H12H12M*. In contrast, dynamic undersampling improved test–set *TSS* for the shorter temporal configurations, *6H6H12M* and *3H3H12M*. These improvements were modest, and the *Recall–Specificity* trade–off differed among the shorter temporal configurations. Therefore, the preferred undersampling strategy appears to be configuration–dependent: static undersampling was favored for the best overall threshold–based classifier, whereas dynamic undersampling remained competitive and provided improvements for the two shorter temporal configurations.

3. Do the predicted probabilities provide meaningful class separation and reliable probabilistic information beyond threshold–dependent classification metrics?

The discrimination and probability–separation analyses showed that all selected models retained strong and broadly comparable threshold–independent discrimination on the independent test set. *ROC–AUC* and *PR–AUC* values were high, ranging from 0.962 to 0.981 and from 0.968 to 0.980, respectively. The dynamic *12H12H12M* model achieved the highest *ROC–AUC*,

while the dynamic *12H12H12M* and dynamic *3H3H12M* models achieved the highest *PR–AUC* values. The predicted–probability distributions also showed a clear separation between positive and negative classes, with positive sequences concentrated near high predicted probabilities and negative sequences concentrated near low predicted probabilities. However, some overlap remained in intermediate probability regions, explaining why final forecasting performance still depends on the selected decision threshold.

The calibration analysis further showed that the predicted probabilities are broadly meaningful, but not uniformly calibrated across all probability ranges. The reliability diagrams indicated generally small average calibration errors, with larger deviations mainly in intermediate probability bins where fewer samples were available. Dynamic undersampling achieved lower *Brier score* for three of the four temporal configurations, whereas static undersampling obtained the lowest *Brier score* for the *24H24H36M* temporal configuration. In terms of *ECE*, dynamic undersampling produced a narrower range of values, although the lowest *ECE* was obtained by the static *24H24H36M* model. The *MCE* values showed a similar mixed behavior, with dynamic undersampling reducing the *MCE* for the *12H12H12M* and *6H6H12M* temporal configurations, whereas the static *24H24H36M* model achieved the lowest overall *MCE*.

Taken together, the results identify the static *12H12H12M* model as the best finalized classifier according to test–set *TSS*. This model reached the highest test–set skill among the evaluated configurations, with $TSS = 0.892$. However, the differences observed in threshold–independent discrimination and calibration suggest metric–dependent tendencies, and do not support a strong conclusion that either undersampling strategy is uniformly superior in terms of probabilistic reliability.

5.5 Conclusion

This chapter presented the sequence–based flare–forecasting phase of the framework developed in this thesis. Building on the representations learned in Chapter 4, the model uses sequences of encoded LoS magnetograms to predict whether an AR will produce at least one M– or X–class flare during a future forecasting window.

This representation–based formulation follows a broader idea in deep learning: high–dimensional inputs can first be mapped into compact latent spaces before being used for downstream prediction. Such representations can reduce the computational burden of working directly with raw data and may provide a feature space in which task–relevant structure is more explicitly organized (Bengio, Courville, and Vincent, 2013; Hinton and Salakhutdinov, 2006). This idea underlies several modern architectures, including vector quantized generative adversarial networks (VQGANs; Esser, Rombach, and Ommer, 2021) and latent diffusion models (LDMs; Rombach et al., 2022), which perform generation in the latent space of a pretrained autoencoder rather than directly in pixel space.

Related ideas have also appeared in flare forecasting, especially in approaches that use image observations directly rather than relying only on manually defined scalar parameters. One closely related example is Chen et al. (2019), who used an autoencoder to extract features directly from SDO/HMI vector magnetograms, applied marginal screening to remove redundant or weakly relevant features, and then used the selected features in an LSTM–based time–series classifier. More recently, Riggi et al. (2025) explored foundation transformer models for flare forecasting across three separate modalities: single magnetograms, magnetogram sequences, and GOES soft X–ray time series.

The approach developed in this thesis differs from these studies in two main respects. First, it does not introduce an additional feature–selection step between representation learning and forecasting, but instead uses the learned magnetogram representations directly in the downstream temporal model. Second, rather than adapting general–purpose pretrained foundation models, both the representation–learning and forecasting components are designed specifically for the LoS–magnetogram sequence–based major–flare forecasting task considered here.

The test–set results show that modeling active–region magnetograms as temporal sequences provides useful predictive information for major flare forecasting, supporting the idea that flare–relevant information is not contained only in the instantaneous magnetic configuration, but also in the temporal evolution of the AR. Among all experimental settings tested, the static *12H12H12M* model achieved the highest test–set *TSS*, with $TSS = 0.892$, followed by the dynamic *12H12H12M* model, with $TSS = 0.856$. In particular, the best overall model maintained a balanced relation between *Recall* and *Specificity*, with $Recall = 0.950$ and $Specificity = 0.942$. The stronger performance of the *12H12H12M* temporal configuration suggests that, among the tested settings, an intermediate observational window provides enough temporal context to capture pre–flare magnetic evolution without introducing excessive variability from longer histories or limiting the model to very short–term changes.

The comparison between static and dynamic undersampling showed that neither strategy was uniformly superior across all temporal configurations. Static undersampling produced the best overall threshold–based classifier and performed better for the longer sequence configurations. Dynamic undersampling remained competitive and improved test–set *TSS* for the shorter configurations. This suggests that the benefit of resampling the majority class at every epoch depends on the temporal structure of the input data.

The threshold–independent analyses showed that all finalized models retained strong discriminative ability on the independent test set. The predicted–probability distributions were generally well separated, although some overlap remained in intermediate–probability regions. This overlap helps explain why the final classification performance depends on the selected decision threshold. The calibration analysis further showed that the predicted probabilities were broadly meaningful, but not uniformly calibrated across all probability ranges. It is worth noting that the best model according to

TSS was not the best calibrated model under any of the calibration metrics. These calibration results correspond to the balanced test set and should not be interpreted as calibration under operational flare prevalence.

Several aspects of the study could be further improved in future work. First, the encoder was kept frozen during forecasting, which allowed the spatial and temporal components of the framework to be separated and reduced the computational cost of the forecasting stage by avoiding further optimization of the encoder parameters. However, this choice also prevented the learned representation from adapting directly to the flare–prediction objective. Future work could compare the frozen encoder with end-to-end training or fine-tuning.

Second, although the models achieved strong forecasting performance, a direct comparison with a single–frame baseline or temporally shuffled sequences is needed to support the interpretation that temporal evolution provides flare-relevant information beyond the instantaneous magnetic configuration. Another limitation of the temporal formulation is that the effect of the observational–window length is not fully separable from the effect of the forecasting–window length, since each configuration uses paired observational and forecasting windows of the same duration. As a result, the present study cannot determine whether performance differences are driven primarily by the amount of past temporal context available to the model or by the length of the future interval over which flares are predicted. This is further complicated by the fact that the 24 h configuration also uses a coarser temporal cadence than the shorter configurations. A more systematic analysis could separate these effects by performing an observational–window sensitivity study, in which the forecasting window is fixed while the observational–window length is varied, and a forecasting–window sensitivity study, in which the observational window is fixed while the forecasting–window length is varied.

Third, the labeling strategy adopted in this chapter introduces an additional limitation. The negative class contains only sequences that remain completely flare–quiet during both the observational and forecasting windows, thereby excluding more challenging operationally relevant situations. The reported results should therefore be interpreted in the context of this controlled discrimination problem rather than as an evaluation of the full operational forecasting problem.

Fourth, additional interpretability analyses could be used to identify which spatial structures and time steps contribute most strongly to the model predictions. Finally, since the calibration results indicate that classification skill and probabilistic reliability do not always coincide, future work could investigate post–hoc calibration methods and alternative imbalance–handling strategies.

The results presented in this chapter support the central idea of the thesis: compact representations learned from magnetogram reconstruction appear to preserve physically relevant information and can serve as useful inputs for downstream forecasting of major flares. In particular, temporally organized sequences of these representations provide useful information for

major–flare forecasting.

Chapter 6

Phase III: Multimodal Spatio–Temporal Flare Forecasting

"All things by immortal power,
Near or far,
Hiddenly
To each other linkèd are,
That thou canst not stir a flower
Without troubling of a star;"

Francis Thompson
The Mistress of Vision

6.1 Introduction

Chapter 5 showed that compact learned representations of line–of–sight (LoS) magnetograms can be used as inputs for sequence–based flare forecasting models. In that phase, the forecasting model relied exclusively on temporal sequences of encoded LoS magnetograms, demonstrating that the temporal organization of learned magnetic–field representations contains useful information for major flare forecasting. However, the magnetic–field information contained in LoS magnetograms provides only one view of the preflare evolution of an active region (AR).

Solar flares are generally understood to result from the release of magnetic energy stored in non–potential coronal magnetic fields through magnetic reconnection (Priest and Forbes, 2002; Forbes, 2000; Schrijver, 2009). Although this energy reservoir is ultimately rooted in the photospheric magnetic field, flare initiation and preflare evolution involve the coupled response of multiple layers of the solar atmosphere. Therefore, combining observations that sample different atmospheric layers of the solar atmosphere can provide a more complete description of the conditions preceding flare occurrence.

The aim of this chapter is to extend the sequence–based forecasting framework developed in the previous phase from a single–modality setting to a multimodal, multiwavelength one. Instead of using only LoS magnetograms, the model is provided with temporally aligned observations from Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) and SDO/Atmospheric Imaging Assembly (AIA), including LoS magnetograms, continuum intensity images, and extreme–ultraviolet (EUV) images in the 193 Å and 304 Å channels, thereby providing a broader view of the solar atmosphere. These image data are complemented by two scalar physical parameters derived from the Space–weather HMI Active Region Patches (SHARP) data product: the total unsigned magnetic flux and the total unsigned current helicity (Scherrer et al., 2012; Lemen et al., 2012; Bobra et al., 2014). In this way, the model combines local spatial information extracted directly from image data with global physical descriptors related to the size, strength and nonpotentiality of the active–region magnetic field.

The forecasting task is formulated as a binary classification problem in which the model predicts whether an AR will produce a flare of class C5.0 or above within the next 2 h, using the previous 36 min of observations. Compared with the previous chapter, this formulation shifts the focus from forecasting major M– and X–class flares over longer temporal windows to a shorter–term forecasting problem that also includes stronger C–class events. This choice allows the model to be evaluated in a near–term prediction setting, where rapidly evolving photospheric and coronal signatures may provide useful information about imminent flare occurrence.

The proposed architecture follows the same general principle adopted in the previous phase: image observations are first mapped into low–dimensional latent feature vectors, and the resulting sequence of feature vectors is then processed by a recurrent classifier. In this chapter, however, the image encoder is trained on multiwavelength image inputs, and the resulting latent representations are fused with SHARP–derived physical parameters before being passed to a recurrent classifier. This design allows the model to learn spatio–temporal patterns from the image sequence while also incorporating physically motivated contextual information.

The chapter is organized as follows. Section 6.2 describes the construction of the multimodal dataset, including the data sources, sequence extraction and labeling procedure, train/validation/test split, and preprocessing. Section 6.3 introduces the forecasting architecture and summarizes the evaluation tools used in this phase. Section 6.4 presents the performance of the three model configurations and discusses the contribution of each input modality, and provides a summary of the key findings. Finally, Section 6.5 discusses the broader implications of these results and outlines the main conclusions of this phase.

6.2 Data

6.2.1 Data Sources

For the study carried out in this chapter, we combined multiwavelength image observations with SHARP–derived physical parameters to construct a multimodal representation of solar ARs. In particular, we used full–disk observations from SDO/HMI and SDO/AIA in their native charge–coupled device (CCD) coordinates, together with the geometrical information contained in the SHARP data series, to identify and extract ARs from the full–disk SDO data. The combined use of these observations and parameters yields a physically consistent and observationally rich dataset suitable for data–driven flare forecasting.

From HMI, we used LoS magnetograms and continuum intensity images. As already discussed in previous chapters, LoS magnetograms provide information on the photospheric magnetic–field distribution, including polarity structure, flux emergence, and flux cancellation. Continuum images provide complementary information on sunspot morphology, area, and temporal evolution, which are observational signatures of the underlying magnetic structure of the AR.

From AIA, we used EUV observations in the 193 Å and 304 Å channels. These two channels were selected because they provide complementary information on the thermal and atmospheric response of ARs during the preflare phase. The 304 Å channel (He II; $\log T \approx 4.7$) samples chromospheric and transition region plasma where filaments, prominences, and early ribbon brightenings originate. The 193 Å channel (Fe XII/XXIV; $\log T \approx 6.2$–7.3) is sensitive to both warm coronal structures and hot flare plasma. Together, these two bands capture the principal atmospheric layers in which flare precursors may manifest, spanning the low solar atmosphere and the corona, and complementing the photospheric information provided by LoS magnetograms and continuum images.

Other EUV bands, such as 171 and 211 Å, have substantial temperature–response overlap with 193 Å (Lemen et al., 2012, Figure 13), and therefore provide less independent diagnostic information. The hot flare channels, 94 and 131 Å, respond most strongly during the impulsive and post–impulsive phases rather than during the preflare evolution relevant to this work (Nishimoto et al., 2021). Moreover, 131 Å includes a significant continuum contribution in ARs, while 211 Å includes a significant continuum contribution in flaring ARs (O'dwyer et al., 2010; Milligan and McElroy, 2013). For these reasons, we restricted the AIA input to the 193 Å and 304 Å channels.

We also selected two photospheric parameters from the SHARP data product that quantify the size, strength and nonpotentiality of AR magnetic fields: the total unsigned magnetic flux (Φ_{tot}) and the total unsigned current helicity ($h_{c,\text{tot}}$), as defined in Bobra et al. (2014). These parameters are derived from vector magnetograms in Cylindrical Equal–Area (CEA) coordinates and are available at a temporal cadence of 12 min.

The total unsigned magnetic flux is defined as in Equation 6.1:

$$\Phi_{\text{tot}} = \sum |B_z| \cdot \Delta A, \tag{6.1}$$

where B_z is the vertical component of the magnetic field, and ΔA is the area of a single pixel. This parameter reflects the magnitude of magnetic flux emerging through the photosphere and serves as a proxy for the total magnetic energy stored in the AR.

The total unsigned current helicity is defined in Equation 6.2:

$$h_{c,\text{tot}} = \sum |B_z \cdot J_z| \cdot \Delta A, \tag{6.2}$$

with the vertical current density J_z given by Equation 6.3:

$$J_z = \frac{1}{\mu_0} \left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y} \right). \tag{6.3}$$

Here, B_x and B_y are the horizontal components of the magnetic field, and μ_0 is the magnetic permeability of free space. The current helicity quantifies the degree of shear and twist in the magnetic field and thus serves as an indicator of magnetic nonpotentiality and energy buildup.

To construct the dataset, we used a list of flare events of class C5.0 and above compiled by Li et al. (2021) from the Geostationary Operational Environmental Satellite (GOES) flare catalog. The list covers the period from 2010 to June 2019, corresponding to most of solar cycle 24, and includes only events that occurred within $\pm 45^\circ$ longitude of the solar central meridian. This restriction reduces the impact of projection effects and center–to–limb–dependent noise. Each event is associated with a National Oceanic and Atmospheric Administration (NOAA) AR number, which is used to link the flare event to the corresponding HMI Active Region Patch (HARP)/SHARP records when possible. In total, 719 events were identified, comprising 21 X–class, 301 M–class, and 397 C–class flares.

6.2.2 Sequence Extraction and Labeling

To construct the positive class for training our binary classification model, we used the flare event list introduced in Section 6.2.1, compiled by Li et al. (2021). Records that could not be matched with at least one HARP were excluded, resulting in 631 flare events, comprising 19 X–class, 270 M–class, and 342 C–class flares. For each flare event, we constructed a single 36 min time window with 12 min steps ending 2 h prior to the flare onset. Using the HARP number and the time range of each sequence, we retrieved the corresponding SHARP data, from which we extracted time series of the two selected physical parameters, Φ_{tot} and $h_{c,\text{tot}}$. We also downloaded the associated full–disk observations — HMI LoS magnetograms, continuum intensity images, and AIA EUV images in the 193 and 304 Å channels — selecting, in each case, the image closest in time to the corresponding reference within the sequence. Thus, unlike the sliding–window procedure adopted in Chapter 5, each flare event contributed at most one positive sequence.

The choice of a 36 min observational window is intended to capture the rapid, short–term evolution that may precede an eruptive event. Observational and numerical studies indicate that preflare signatures — such as localized flux cancellation, increases in magnetic shear, and the early expansion or brightening of coronal loops — can develop on timescales of minutes to tens of minutes prior to flare onset (e.g., Fletcher et al., 2011; Benz, 2017; Toriumi and Wang, 2019). Other studies demonstrate that once a flare begins, the magnetic field can undergo abrupt restructuring on similarly short timescales (e.g., Sudol and Harvey, 2005; Wang, 2006; Sun et al., 2012). This short observational window is also logically matched to our 2 h forecast horizon: for near– term prediction, the most recent rapid changes in the magnetic and coronal environment are generally more informative than conditions several hours or days earlier.

Since HMI and AIA differ in their world coordinate system (WCS; Greisen and Calabretta, 2002) definitions, spatial alignment was also required. We used AIA 304 Å data as a spatial reference, and all other data were reprojected onto its WCS grid using solar–aware coordinate transformations. This step ensures that all data are mapped to the same physical coordinates on the solar disk, providing a coherent, multiwavelength dataset in which photospheric and coronal observations are both temporally and spatially aligned.

To identify the location and extent of ARs on the solar disk and extract consistent cutouts, we relied on SHARP WCS. We adopted a fixed reference size of 512×512 pixels. If the dimensions of the AR bounding box were smaller than this size, a square patch of 512×512 pixels centered on the AR was extracted. If one or both dimensions of the AR were larger than this size, the full extent of the region along that dimension was retained, so that the final cutout size was determined independently for each dimension (Figures 6.1 (a) and (b)).

We excluded very small ARs, defined as regions for which at least one original dimension is smaller than 200 pixels, because such regions would occupy only a small fraction of a 512×512 cutout and would therefore be dominated by background signal. For large ARs, this extraction strategy avoids downscaling or clipping and preserves the complete spatial structure. For smaller regions that still exceed 200×200 pixels, enforcing a minimum size ensures that the extracted patches maintain an adequate signal–to–background ratio.

Finally, to ensure data quality, we discarded any sequences containing at least one frame that crossed the solar limb (Figure 6.1 (b)). This approach naturally results in cutouts from different ARs having varying dimensions. Further details on handling variability in image size across different sequences during training are provided in Section 6.2.4.

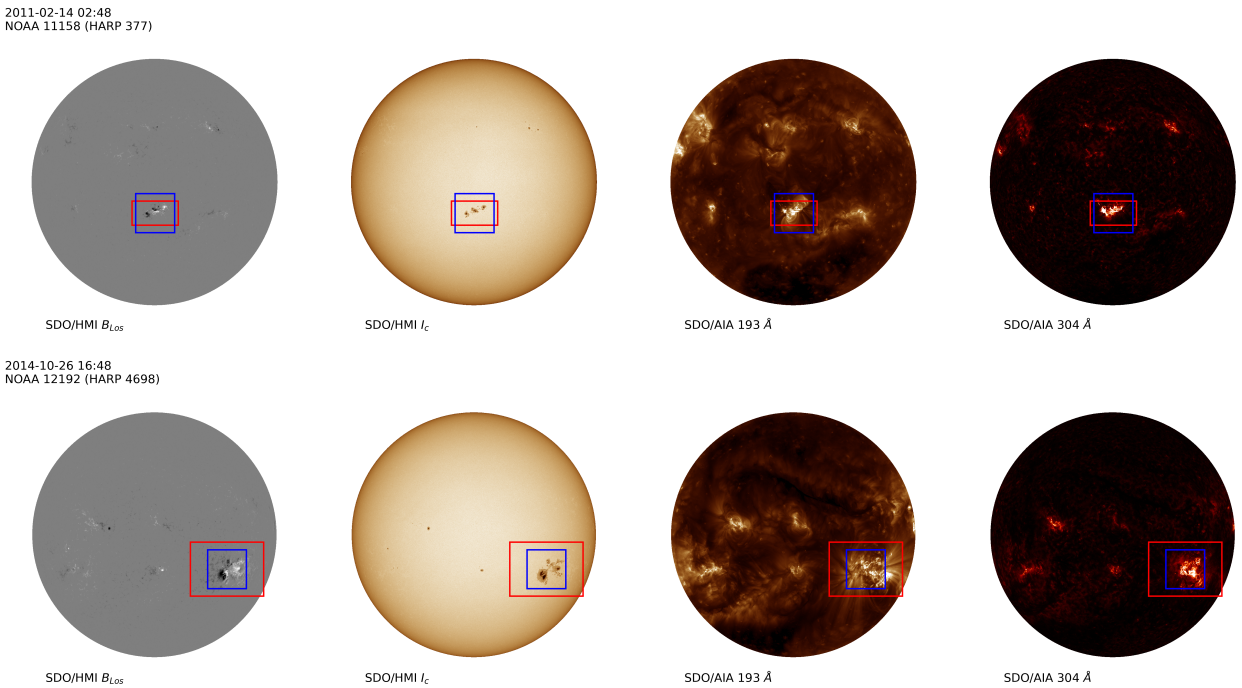


FIGURE 6.1: Examples of active region extraction and quality control. Each panel shows, from left to right, SDO/HMI LoS magnetograms, SDO/HMI continuum intensity images, and SDO/AIA EUV images in the 193 and 304 Å channels, all reprojected to the AIA 304 Å WCS and taken 2 h prior to the flare peak time. Red boxes delineate the full extent of the active region extracted from the SHARP data product, while blue boxes indicate the reference dimensions of the adopted cutouts. (a) NOAA 11158 (HARP 377): HARP original bounding–box size was 377×744 pixels, exceeding 512 pixels in one dimension but not the other. Consequently, the final cutout extent along each axis was defined as the greater of 512 pixels and the original dimension, resulting in a cutout of 512×744 pixels. (b) NOAA 12192 (HARP 4698): the bounding box enclosing the active region crossed the solar limb within the look–back window; consequently, the corresponding sequence was excluded from the dataset.

To construct the negative class, we applied a systematic selection strategy aimed at avoiding contamination from flare–related activity. First, we compiled a list of all ARs present in the SHARP data, extending the observational period from 2010 to 2024, covering most of solar cycle 24 and the early phase of solar cycle 25, to ensure a sufficiently large pool of candidate ARs from which to select our final negative examples. We then excluded all ARs included in the positive class. For each remaining AR, we examined its SHARP records to identify valid data windows that met the following criteria: (i) each record is located within $\pm 45^{\circ}$ in both Stonyhurst longitude and latitude, (ii) the AR remains flare–quiet for at least 3 h before and after the end of the sequence, and (iii) all records are of good quality (i.e., all QUALITY flags are below 65536 [1]).

This approach ensures a clean separation between flare and nonflare examples for training. The final dataset includes 562 flare sequences and 644 nonflare sequences. The class distribution is summarized in Table 6.1.

Class	Total	C–class ($\geq$ C5.0)	M–class	X–class
Flare	562	303	242	17
No Flare	644	–	–	–

TABLE 6.1: Summary of the dataset composition grouped by GOES flare class. The flare category includes events above the C5.0 threshold, with each valid flare event contributing a single observational sequence. The no–flare category denotes samples with no associated flaring activity within the prediction window.

Our dataset is structured as multiwavelength — HMI LoS magnetograms, continuum intensity, and AIA EUV channels —, multimodal — raw images and scalar physical parameters —, and multistep — time series—, enabling the predictive model to learn from the evolving physical conditions within ARs and identify spatiotemporal patterns that may serve as flare precursors. For the image data, no explicit feature engineering is applied; instead, the model operates directly on raw images, allowing it to autonomously learn meaningful representations. In parallel, physically motivated scalar parameters are incorporated to provide complementary global context. This design also allows the model to assign relative importance to each modality during classification, deciding whether to focus on local features extracted from images or global parameters provided as input.

6.2.3 Train/Validation/Test Split

To ensure fair evaluation the dataset was divided into nonoverlapping training, validation, and test sets. As in Chapters 4 and 5 the split was performed at the AR level. For the positive class, corresponding to flaring ARs,

[1] http://jsoc.stanford.edu/doc/data/hmi/Quality_Bits/QUALITY.txt

we randomly divided the list of ARs into three mutually exclusive groups and assigned all sequences associated with a given HARP to the same group, targeting approximate proportions of 70%/15%/15% in the number of sequences assigned to each of the training, validation, and test sets. The same procedure is applied independently to the negative class, corresponding to quiet ARs.

After splitting the positive and negative samples, we combined the corresponding subsets to form the final training, validation, and test sets. This strategy ensures that each set contains a balanced mix of flare and quiet sequences, with no AR appearing in more than one subset. Assigning all sequences from any given AR exclusively to a single split prevents the model from encountering temporally or spatially related data across sets and ensures that performance metrics reflect the model's ability to generalize to unseen ARs. An AR-level split provides a more realistic evaluation of model generalization by avoiding information leakage between different sets. However, the class distribution itself is balanced by construction and therefore does not reproduce the strong class imbalance characteristic of the flare forecasting problem.

6.2.4 Preprocessing

To normalize the input data, we compute the global minimum and maximum values for each image channel independently, using only the training set. Each image is then scaled to the range $[-1, 1]$ using its corresponding channel statistics. A similar procedure is applied to the physical parameters, where each parameter is individually scaled to the same range $[-1, 1]$ using the global minimum and maximum values computed from the training samples. Following the scaling strategy discussed in Section 4.2.4, these training–set statistics are then applied to the validation and test sets to maintain consistency across splits.

To address missing values, we apply channel–wise median imputation. For each image, missing values are replaced with the median of the corresponding channel within that image. For the physical parameters, missing values are replaced with the median of the corresponding parameter, computed across all training samples. This approach helps preserve the central tendency of the data without introducing significant bias.

As described in Section 6.2.2, the extracted cutouts vary in size depending on the dimensions of each AR bounding box. Since the model requires inputs with fixed spatial dimensions, a strategy to standardize input size is required. Unlike the data used in Chapters 4 and 5, where cutouts from the SHARP data series were used, here we extracted the AR cutouts from full–disk observations with a minimum size of 512×512 pixels. Therefore, no padding is required in this case, and the standardization can be performed through cropping alone.

As in the previous chapters, we adopt random cropping during training to standardize the input dimensions to the reference size of 512×512 pixels.

By varying the crop location at each epoch, the model is exposed to different subregions of the AR during training, allowing it to gradually learn from the full spatial context without requiring explicit resizing or padding. This method serves both as a standardization step and as a regularization strategy, making the model less prone to overfitting by creating a virtually larger dataset.

6.3 Methodology

6.3.1 Model

The proposed deep learning framework is designed to predict whether a solar flare of class $\geq$ C5.0 will occur within the next 2 h, based on the previous 36 min of observations. The framework integrates multiwavelength image sequences and scalar physical parameters through a hybrid architecture that captures both spatial–temporal patterns and contextual information. It is composed of a convolutional encoder that extracts features from images, a fusion mechanism that combines image–derived features with the corresponding physically derived parameters, a recurrent network that models the temporal evolution of the fused representation, and a fully connected classification network that maps the learned temporal features to a probabilistic prediction of flare occurrence. A conceptual sketch of the model architecture is illustrated in Figure 6.2.

The input to the model consists of sequences of four–channel images, including LoS magnetograms, continuum intensity images, and observations from the 193 Å and 304 Å channels. These image sequences are complemented by time series of two scalar physical parameters: the unsigned magnetic flux and the total unsigned current helicity, both obtained from SHARP data products (see Section 6.2.1). Thus, each image in the sequence is temporally matched with the corresponding values of the scalar parameters.

The training strategy was divided into two parts. The first part consisted of training an autoencoder on individual four–channel images. An autoencoder is a type of artificial neural network trained to learn compressed representations of data. It is composed of an encoder, which transforms the input data into a low–dimensional latent space, and a decoder that reconstructs the input from this compact representation. This process enables the model to extract essential features useful for downstream tasks.

In our implementation, the encoder is composed of five convolutional blocks; each block includes a convolutional layer that also performs spatial downsampling, followed by a rectified linear unit (ReLU) activation function, which progressively transforms a single four–channel 512×512 input into 512–channel 16×16 feature maps. The decoder mirrors this architecture in reverse: each decoding block includes a transposed convolutional layer that also performs spatial upsampling (Dumoulin and Visin, 2016), followed by a ReLU activation function, except in the final layer, where we used a hyperbolic tangent (*tanh*) activation function to constrain the output range to

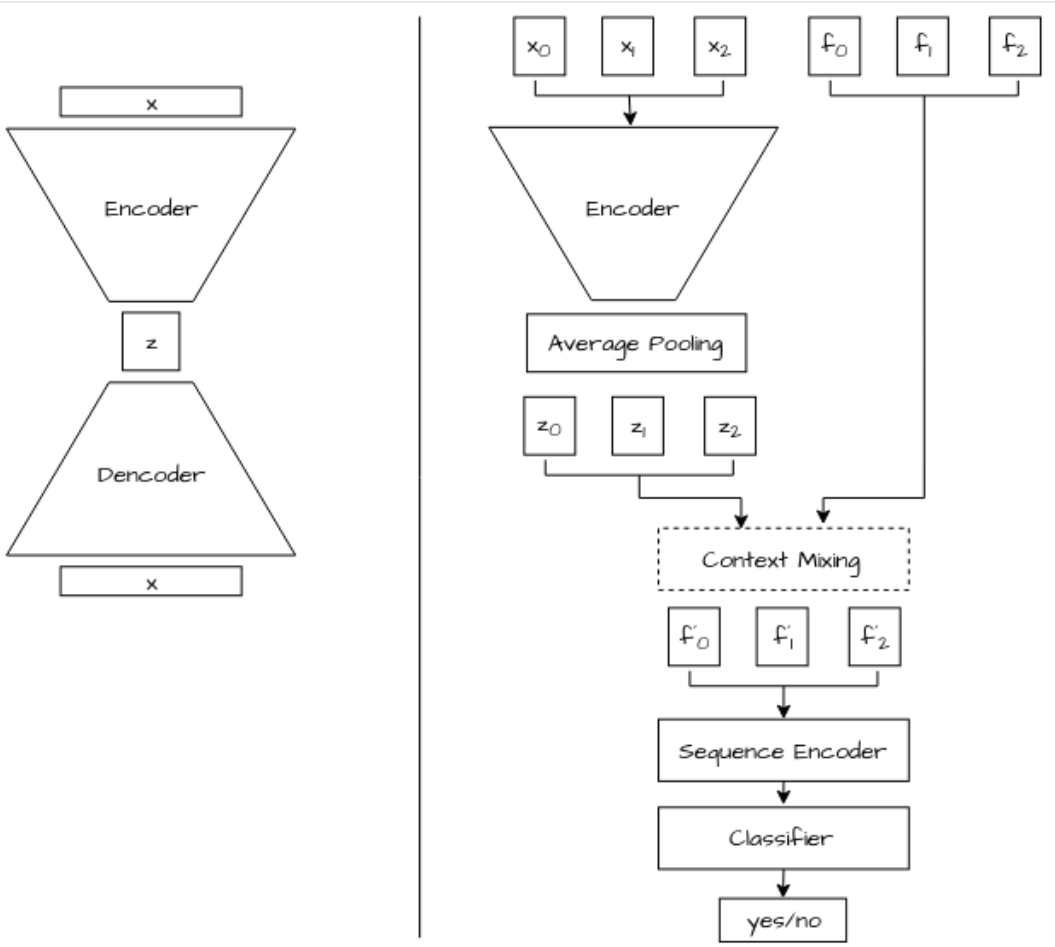


FIGURE 6.2: Overview of the proposed solar flare prediction framework. Left: a convolutional autoencoder is trained on individual multi–wavelength solar images (x) to learn latent representations (z), enabling unsupervised feature extraction. Right: a supervised prediction pipeline integrates learned image representation and physical features. Latent representations (z_0, z_1, z_2) are obtained by applying the encoder independently to each image in the sequence, while hand–engineered physical parameters (f_0, f_1, f_2) undergo a linear transformation to match the dimensionality of the learned features, allowing the two modalities to be summed and fused into a context–aware representation. The fused feature sequence is then processed by a recurrent sequence encoder, followed by a classifier that outputs the predicted probability of a flare event.

$[-1, 1]$, which corresponds to the scaling range of the input data. Following training, the decoder is discarded, and the encoder is retained with fixed parameters and used as a feature extractor for subsequent temporal modeling. As in the previous phase, this training strategy was adopted to maintain a separation between representation learning and forecasting, while also reducing the computational cost of the forecasting stage by limiting the number of trainable parameters. To obtain a compact representation for integration with scalar parameters, we applied global average pooling across the spatial dimensions of the encoder output, reducing each $512 \times 16 \times 16$ feature map to a single 512–dimensional feature vector.

To integrate the scalar physical parameters with the image–based representation, a context mixing module is introduced. The two–dimensional vector of physical parameters is projected into 512 dimensions using a fully connected layer. The projected context vector is then combined with the learned representation of the image data through element–wise addition, enabling joint modeling of image and contextual physical information. We also experimented with alternative fusion strategies, including element–wise multiplication, feature concatenation, and attention–based mechanisms (Vaswani et al., 2017), and observed comparable performance across these methods.

The resulting fused vectors, computed at each time step, are fed into the sequence encoder based on a long short–term memory (LSTM) network, which models temporal dependencies across the input sequence. The final output of the LSTM is passed through a classification module consisting of two fully connected layers with ReLU and a final sigmoid activation. The sigmoid constrains the output to the $[0, 1]$ range, allowing it to be interpreted as the model's predicted probability of solar flare occurrence for a given input sequence. This combined architecture enables the model to learn spatiotemporal and contextual relationships relevant to the prediction of solar flares.

The autoencoder was trained using the Adam optimizer with a learning rate of 0.001 and mean squared error loss over 300 epochs, with a batch size of 32. Details of the implementation are summarized in Table 6.2.

For the sequence modeling stage, we experimented with the Adam and AdamW optimizers (Kingma and Ba, 2014; Loshchilov and Hutter, 2017), varying the learning rates over $\{10^{-5}, 10^{-4}, 10^{-3}\}$, LSTM hidden sizes over $\{8, 32, 64, 128\}$, and layer depths over $\{1, 2, 3, 4\}$. For AdamW, weight decay values of $\{10^{-4}, 10^{-3}, 10^{-2}\}$ were also explored. Binary cross–entropy was used as the classification loss, and the best model was selected based on the minimum validation loss. The results reported in Section 6.4 correspond to the architecture summarized in Table 6.3, trained using the Adam optimizer with a learning rate of 0.0001 for 100 epochs.

6.3.2 Evaluation Tools

The evaluation protocol adopted in this chapter is the same as that introduced in Section 5.3.2. The flare–forecasting problem is treated as a binary classification task, so the model output is interpreted as the predicted probability of flare occurrence and is converted into a class label by applying a

Component	Encoder	Decoder
Layer type	Convolution	Transposed convolution
Number of blocks	$\times 5$	$\times 5$
Channels	$4 \to 64 \to 128$ $\to 256 \to 512 \to 512$	$512 \to 512 \to 256$ $\to 128 \to 64 \to 4$
Kernel / stride / padding	$3 \times 3/2 \times 2/1$	$4 \times 4/2 \times 2/1$
Activation	ReLU $\times 5$	ReLU $\times 4$, *tanh* final

TABLE 6.2: Summary of the convolutional autoencoder architecture used to compress individual four–channel input images into compact latent representations. The encoder consists of five convolutional blocks with increasing channel depth, while the decoder mirrors this structure using transposed convolutions to reconstruct the input image. Activation functions include ReLU for intermediate layers and *tanh* at the output of the decoder.

Component	Layer type	Details
Context mixing	Linear	$2 \to 512$
LSTM	LSTM	Input size $= 512$, hidden size $= 32$ layers $= 2$
Classifier	Linear	$32 \to 32$
	ReLU	Activation
	Linear	$32 \to 1$
	Sigmoid	Output probability $(0, 1)$

TABLE 6.3: Architecture of the sequence classifier used to predict flare occurrence from latent representations. The model consists of a context mixing layer that projects the physical parameters into the same dimensional space as the latent image representations, followed by a two–layer LSTM for sequence processing and a fully connected classifier. A final sigmoid activation outputs the probability of a flare event within the prediction window.

decision threshold. Unless otherwise stated, the default threshold $\tau = 0.5$ is used. Therefore, a sequence is classified as flare–productive when the predicted probability is greater than or equal to τ, and as flare–quiet otherwise.

Model performance is assessed using the confusion–matrix–based metrics defined in Equations 5.7–5.14. In particular, we report the number of true positives (*TP*), false positives (*FP*), true negatives (*TN*), and false negatives (*FN*), together with *Accuracy*, *Precision*, *Recall*, *F1–score*, and *TSS*. These metrics provide complementary information on the behavior of the classifier: *Accuracy* measures the overall fraction of correct classifications; *Precision* quantifies the reliability of positive flare forecasts; *Recall* measures the ability to detect flare–producing cases; *F1–score* summarizes the balance between *Precision* and *Recall*; and *TSS* measures the difference between the *TPR*, also known as *Recall* or *Sensitivity*, and the *FPR*, providing a class–balance–aware estimate of forecasting skill.

In addition to these scalar metrics, model behavior is analyzed using the same diagnostic plots introduced in Section 5.3.2. Predicted–probability distributions are used to assess the separation between flare–positive and flare–negative samples. The receiver operating characteristic (ROC) curve and the corresponding *ROC–AUC* values summarize the threshold–independent discriminative capability of each model. Calibration curves compare predicted probabilities with observed flare frequencies, providing a qualitative assessment of whether the model outputs can be interpreted as reliable probability estimates. Finally, *Precision*, *Recall*, and *F1–score* are analyzed as functions of the decision threshold in order to evaluate the sensitivity of each model to threshold selection.

6.4 Results

We present an empirical evaluation of the performance of our solar flare prediction framework, which consists of a deep learning model trained on multimodal observational data to classify solar ARs as either flare–imminent or flare–quiet. Three model configurations are evaluated, each sharing the same underlying architecture but differing in their input modalities and optimized hyperparameters. The baseline model, referred to as model M, is trained exclusively on LoS magnetograms. Model MCA extends this configuration by incorporating continuum intensity images and EUV observations. The most comprehensive model, model MCAF, further integrates scalar physical parameters derived from SHARP vector magnetogram data, which encode magnetic field properties known to correlate with flaring activity.

Our analysis is designed to address three primary questions. (1) Does the addition of continuum and EUV imagery lead to measurable improvements in predictive performance relative to the LoS magnetogram–only baseline? (2) Does incorporating SHARP–derived physical parameters, which reflect expert–driven diagnostics, enhance classification performance beyond what is achieved with imagery alone? (3) To what extent do each model's predicted probabilities align with empirical flare rates across varying thresholds, thereby quantifying calibration quality?

To answer these questions, we conduct a controlled comparison of the three models on a balanced dataset using consistent training procedures and evaluation protocols. All performance metrics reported in this section are computed on the balanced test set described in Section 6.2.3. In the following sections, we present both quantitative and qualitative analyses that demonstrate improvements in predictive performance and highlight the contributions of each observational modality.

Classifier (Base Model: Only LoS Magnetogram)

The baseline model (model M), trained exclusively on LoS magnetograms, establishes a solid reference point for flare prediction using single–modality input. It achieves an *Accuracy* of 87.71%, correctly classifying the majority of ARs. A *Recall* of 87.80% indicates a good ability to detect flaring events, capturing nearly 9 out of 10 actual flares. The model's *Precision*, at 85.71%, suggests that most predicted flares are correct, with approximately 14% of the predicted flares being false alarms. This balance between detecting true flares and minimizing unnecessary alerts is reflected in an *F1–score* of 0.87. The *TSS*, which accounts for both missed detections and false alarms, is 0.75, indicating that the model performs better than random chance and fairly distinguishes between flare–imminent and quiet regions (see Table 6.4 and 6.5).

The predicted probability distribution (Figure 6.3, first row, left) illustrates how the model separates flare–imminent and flare–quiet ARs (positive and negative class). While there is a visible tendency for flare–imminent regions to receive higher probability scores, the distributions for both classes exhibit substantial overlap. This overlap indicates that the model's discriminative capacity is moderate, with many quiet regions assigned intermediate probabilities and not all flare regions predicted with high confidence. This plot provides a visual counterpart to numerical metrics such as *Accuracy* and *F1–score*, revealing uncertainties that these scores alone do not capture.

The calibration curve (second row, left) compares predicted flare probabilities with the observed flare frequencies and is computed here using two binning strategies: uniform, which guarantees equally spaced bins, and quantile, which guarantees the same number of samples in each bin. The curve shows deviations from the ideal diagonal, indicating slight overconfidence at intermediate and high probability ranges, where the predicted likelihoods exceed the observed flare rates, particularly for higher probabilities (> 0.8). This pattern reinforces insights from the probability distribution: many predictions fall into ambiguous ranges where class separation is weak. Despite these deviations, the overall trend remains monotonic and consistent, suggesting that the model's probability outputs are still informative, though they would benefit from calibration for operational use.

The ROC curve (third row, left) shows good discriminative power, with an *ROC–AUC* of 0.95. Unlike the probability distribution and calibration curve, which reveal limitations in prediction confidence and probability alignment, the ROC curve focuses only on the model's ability to rank–order instances correctly, regardless of probability threshold. Thus, it demonstrates

that the model maintains discriminative capability even in the presence of calibration errors.

The performance curves as a function of threshold (fourth row, left) complement the ROC analysis by highlighting model performance across decision thresholds. *Precision* remains high until very low thresholds (< 0.2), indicating that additional predicted flares are mostly true events. *Recall* remains high up to around a threshold of 0.5, where it stays between 0.85 and 0.90, corresponding to a local *F1–score* maximum that offers the best balances between *Precision* and *Recall* — consistent with the default classification threshold of 0.5. This curve confirms the *Recall*–oriented behavior observed in the ROC and *F1–score* analyses while also showing that thresholds favoring higher *Recall* incur a modest reduction in *Precision*.

Classifier (Base Model + Continuum and EUV)

Model MCA, which incorporates continuum intensity and EUV observations in addition to LoS magnetograms, shows substantial improvements across all key classification metrics. *Accuracy* increases to 90.50%, with a notable rise in *Recall* to 93.90%, indicating that the model captures a greater proportion of true flaring events. *Precision* also improves to 86.52%, resulting in a higher *F1–score* of 0.90. The *TSS* increases to 0.82, demonstrating an enhanced ability to separate flaring from nonflaring regions. These results confirm the benefit of incorporating additional observational data that provide complementary information on thermal and morphological flare precursors.

The predicted probability distribution (Figure 6.3, first row, center) shows improved class separation, with flare–imminent and quiet regions forming more distinct clusters than in the baseline model. The calibration curve (second row, center) exhibits moderate underconfidence at intermediate probabilities, where predicted likelihoods underestimate the true flare rate, but it aligns closely with the ideal diagonal at low and high probabilities, indicating reliable estimates at the extremes. Improvements are further reflected in the ROC curve (third row, center), which shows a higher *ROC–AUC*, and in the performance–versus–threshold curve (fourth row, center), where intermediate thresholds yield optimal *F1–score*. These visual diagnostics reinforce the conclusion that multiwavelength imaging significantly enhances the model's predictive accuracy and reliability.

Classifier (Base Model + Continuum and EUV + Physical Quantities)

Model MCAF, which extends model MCA by incorporating SHARP–derived physical quantities in addition to LoS magnetograms, continuum intensity, and EUV observations, achieves the strongest performance among all configurations. *Accuracy* increases to 92.74%, and *Recall* rises to 97.56%, with only 2 out of 82 flares not predicted by the model. *Precision* rises slightly to 87.91%, yielding a higher *F1–score* of 0.92. *TSS* reaches 0.86, reflecting superior discriminative ability. These observed improvements suggest that domain–specific magnetic diagnostics provide complementary information for flare forecasting.

This enhanced performance is corroborated by the evaluation plots. The predicted probability distribution (Figure 6.3, first row, right) shows even greater class separation compared to model MCA, with flare–imminent regions clustering tightly at high probability values.

The calibration curve (second row, right) demonstrates overall good probability alignment. The uniform–binned curve (blue) presents a sharp dip around midrange probabilities. This dip is consistent with a low–count artifact from equal–width bins and reflects the underlying probability distribution, which is strongly bimodal — most negatives near 0, most positives near 1, with almost no mass in $0.2 - 0.8$. In contrast, the quantile–binned curve (orange), which guarantees an equal number of samples in each bin, tracks the 45° line closely. At low and high predicted probabilities, both binning methods converge toward the diagonal, suggesting that the model is well calibrated in the regions where most predictions lie.

The ROC curve (third row, right) rises steeply toward the upper left corner, achieving an *ROC–AUC* of 0.98. The near–vertical rise at low *FPR* reflects strong *Sensitivity* without substantial compromise in *Specificity*, reflecting robustness in its discriminative performance.

The threshold–dependent performance curves (fourth row, right) show that *Precision*, *Recall*, and *F1–score* remain uniformly high across a broad range of thresholds. *Recall* (orange) remains near–perfect across a wide range of thresholds (< 0.8), only dropping sharply when the threshold approaches 1.0. *Precision* (blue) is stable at high levels for most thresholds, with only a modest decline at very low thresholds (< 0.2), where *FP* are more common. The *F1–score* (green) maintains values close to its maximum over a broad threshold interval ($0.2 - 0.8$), indicating robustness in balancing *Precision* and *Recall*. This stability suggests that the model's performance is not highly sensitive to the exact threshold choice, providing flexibility for operational settings depending on whether higher *Recall* or higher *Precision* is needed.

Model	**TP**	**TN**	**FP**	**FN**
Model M	72	85	12	10
Model MCA	77	85	12	5
Model MCAF	80	86	11	2

TABLE 6.4: Confusion matrix values for all models evaluated in this study. All quantities are computed on the held–out test set.

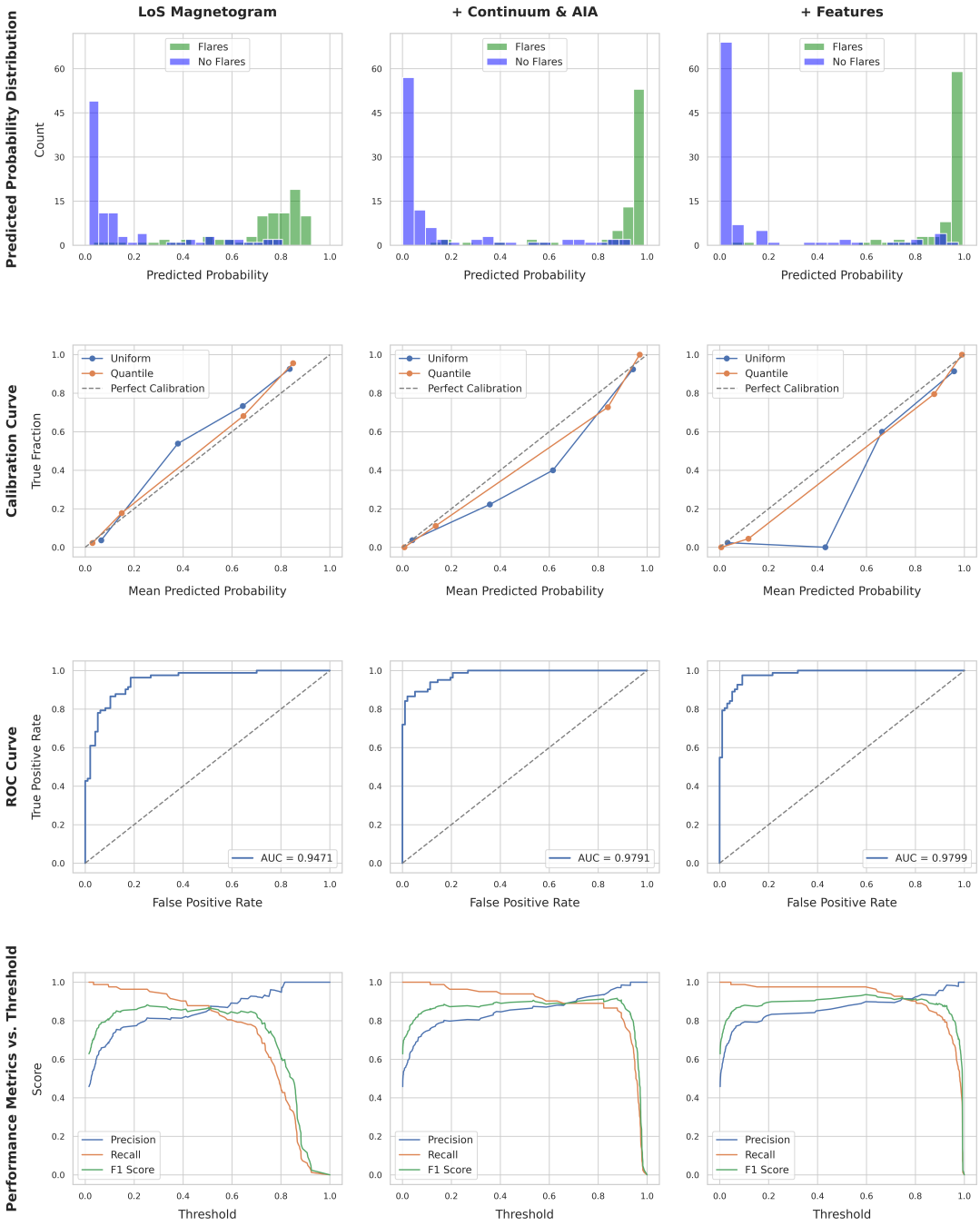


FIGURE 6.3: Classifier performance across different input modalities. Columns represent models trained on: (1) LoS magnetograms only; (2) LoS magnetograms with additional continuum and AIA image channels; and (3) full multimodal input including SHARP physical parameters. Top row: predicted probability distributions for positive and negative samples. Second row: calibration curves comparing predicted probabilities with observed outcome frequencies. Third row: ROC curves and *ROC–AUC* values, showing model discrimination capability. Bottom row: *Precision*, *Recall*, and *F1–score* as functions of the decision threshold. Performance improves consistently as additional data modalities are incorporated.

Model	Accuracy	Precision	Recall	F1–score	TSS
Model M	0.8771	0.8571	0.8780	0.8675	0.7543
Model MCA	0.9050	0.8652	0.9390	0.9006	0.8153
Model MCAF	0.9274	0.8791	0.9756	0.9249	0.8622

TABLE 6.5: Performance metrics for all models evaluated in this study. All quantities are computed on the held–out test set.

6.4.1 Summary of Key Findings

To address the three questions posed at the beginning of this section, we designed experiments to systematically evaluate the benefits of incorporating multiwavelength observations together with physical parameters. Our deep learning framework, while leveraging state–of–the–art architectures, was specifically conceived for this task and is capable of jointly processing a rich, multimodal dataset. In what follows, we summarize the key findings of our analysis in light of the aforementioned questions.

1. Does the addition of continuum and EUV imagery lead to measurable improvements in predictive performance relative to the LoS magnetogram–only baseline?

We find that the inclusion of continuum and EUV channels substantially improves model performance. Comparing model M (Los magnetogram–only) to model MCA (LoS magnetogram + continuum + EUV), we observe a notable increase in performance across all key metrics: *Accuracy* improves from 87.71% to 90.50%, the *F1–score* rises from 0.87 to 0.90, and the *TSS* improves from 0.75 to 0.81. This improvements are corroborated by the ROC curve and the predicted probability distribution, which show a clear improvement in class separation. Furthermore, the performance metrics versus threshold curves demonstrate that the MCA model maintains consistently higher *F1–score* across a wide range of thresholds, indicating greater robustness to threshold selection. These results suggest that photospheric and coronal emissions captured by continuum and EUV imaging provide complementary and predictive information not contained in LoS magnetograms alone.

2. Does incorporating SHARP–derived physical parameters, which reflect expert–driven diagnostics, improve classification performance beyond what is achieved with imagery alone?

The addition of scalar physical parameters — total unsigned magnetic flux and current helicity — is associated with a further improvement in performance. When comparing model MCA to the full multimodal configuration (model MCAF), the *Accuracy* improves from 90.50% to 92.74%, the *F1–score* from 0.90 to 0.92, and the *TSS* from 0.82 to 0.86, indicating strong classification skill. *Recall* increases from 93.90% to 97.56%, corresponding to three additional true positives among the 82 positive test samples; one additional negative sample is correctly classified. Improvements are also evident

in the probability distributions and ROC curve, and the stability of the performance curves across a wide range of thresholds highlights the model's balanced behavior between *Precision* and *Recall*. These results suggest that physically interpretable global features provide valuable complementary information when combined with imaging data. Although the direction of the observed improvement is clear, uncertainty estimates are necessary to assess the robustness and magnitude of this incremental gain.

3. To what extent do each model's predicted probabilities align with empirical flare rates across varying thresholds, thereby quantifying calibration quality?

The predicted probabilities of all three models align well with empirical flare rates, indicating that calibration quality is generally good. At first glance, calibration seems to degrade as *Accuracy* increases. This effect is likely an artifact of deep learning models concentrating predictions near 0 and 1, which results in low counts and less reliable statistics in the mid–probability ranges. When using a quantile–based binning strategy, however, the curves show that all three models are fairly well calibrated, with slight improvements when continuum, EUV, and physical parameters are included. This improvement is particularly evident in the low– and high– probability regions where most predictions lie. Overall, this progression demonstrates that richer input data not only improve classification metrics but also maintain relatively good calibration, which is critical for operational forecasting. Well–calibrated predictions enable decision–makers to assess flare likelihood more reliably and to adjust thresholds according to risk tolerance.

6.5 Conclusions

In this chapter, we extended the sequence–based flare forecasting framework developed in Chapter 5 to a multimodal and multiwavelength setting. The proposed model integrates temporally ordered image observations from SDO/HMI and SDO/AIA with scalar physical parameters derived from SHARP vector magnetograms. In this way, the framework combines spatio–temporal information extracted from solar atmospheric observations with global descriptors of active–region magnetic properties.

The results show that predictive performance improves as additional sources of information are incorporated into the model. The baseline configuration, based only on LoS magnetograms, already achieves good classification performance, confirming that the temporal evolution of photospheric magnetic–field information contains useful predictive information. Adding continuum and EUV imagery further improves the results, indicating that multiwavelength observations provide complementary information about the preflare evolution of ARs. The highest performance is obtained when the SHARP–derived physical parameters are also included. On the balanced test set, this configuration reaches an *Accuracy* of 92.74%, a *Recall* of 97.56%, and a *TSS* of 0.86. Relative to the LoS magnetogram–only model M, the full MCAF configuration correctly classifies eight additional positive samples and one additional negative sample. Relative to model MCA, model MCAF correctly

classifies three additional positive samples and one additional negative sample. This favorable direction suggests a potential complementary contribution from the physical parameters; however, because no uncertainty estimates were obtained, the robustness and magnitude of this incremental improvement cannot be established from the present results. Performing such an analysis would provide clearer empirical support for the benefit of the richer multimodal input configuration.

The diagnostic analyses support this conclusion. The predicted–probability distributions show an increasingly clear separation between flaring and nonflaring sequences as the input configuration becomes richer. The ROC curves analysis confirms the improved threshold–independent discriminative capability of the multimodal models, while the threshold–dependent curves show that the full configuration maintains strong performance over a broad range of decision thresholds. The calibration analysis indicates that the predicted probabilities are generally informative, although deviations from perfect calibration remain, particularly at intermediate probability values. This suggests that post hoc calibration could be useful in applications where the probabilistic interpretation of the model output is important.

Despite these promising results, some limitations remain. First, in this work, we focus on predicting flaring versus quiet ARs. Our sequence extraction strategy, in particular the construction of the negative class (no flaring), excludes sequences with decaying flare activity as well as sequences leading to flares below the chosen threshold ($\leq$ C5.0). Although our sequence extraction and labeling approach is similar to those used in other sequence–based studies (e.g., Sun et al., 2022; Angryk et al., 2020), our findings may not be immediately extended to other task definitions. For example, this is the case in operational settings, where flare–quiet regions that evolve from previously flare–active states are of interest. One possible solution could be to include the flaring history of the AR as an additional input, similar to the approach taken in Jonas et al. (2018). Such information would provide the necessary context for the model to distinguish truly quiet regions from those that are flare–capable.

Second, we intentionally constructed a balanced dataset in order to avoid bias toward the majority nonflaring class. While this choice is useful for evaluating the intrinsic discriminative capability of the model, real solar–flare forecasting is characterized by strong class imbalance. The reported performance metrics should therefore be interpreted within this controlled evaluation setting and not as estimates of performance under operational event rates. Future work should therefore assess the same multimodal framework under more realistic event rates and investigate training strategies designed for imbalanced datasets, including cost–sensitive learning or score–oriented loss functions (Spelmen and Porkodi, 2018; Ahmadzadeh et al., 2021; Marchetti et al., 2022).

Another limitation related to dataset construction concerns its temporal composition. Positive sequences are derived from the Li et al. (2021) flare–event list, which extends from 2010 to June 2019, whereas negative sequences are sampled from observations extending to 2024, with a substantial number

occurring during Solar Cycle 25. Consequently, an epoch–dependent contribution to class separation cannot be excluded. This is particularly relevant for the AIA 304 Å channel, which is known to be affected by time–dependent sensitivity degradation and calibration uncertainties (Boerner et al., 2014; Zwaard et al., 2021). A temporally matched negative–sampling strategy or an explicit degradation correction would therefore help quantify the extent of this effect.

Finally, the scale of the dataset remains an important practical constraint. Our dataset is relatively small in terms of the number of samples, but each sample consists of a temporal sequence of high–resolution multichannel images together with scalar physical parameters. Considering more samples would require greater computational resources, both for training and for dataset preprocessing and storage. In this work, we partially address this issue by randomly cropping the images at each epoch, virtually increasing the size of the dataset. Expanding the dataset remains the most effective solution, ideally by incorporating observations from multiple solar cycles. The same practical considerations apply when using larger image sizes or longer temporal sequences are considered, as this would further increase memory requirements and training time and impose additional constraints on batch size and model architecture.

Chapter 7

Physical Interpretation of Multimodal Solar Flare Forecasting

"For now we see only a reflection as in a mirror; then we shall see face to face. "

1 Cor 13:12

7.1 Introduction

The previous chapter developed a multimodal and multiwavelength flare–forecasting model that combines image sequences from Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) and SDO/Atmospheric Imaging Assembly (AIA) with scalar physical parameters derived from vector magnetograms. The results showed that incorporating complementary sources of information improves forecasting performance, particularly when the model combines spatio–temporal image information with global magnetic descriptors. However, predictive performance alone does not explain how the model forms its decisions, which input sources contribute most strongly to each prediction, or whether the learned decision patterns are consistent with physically meaningful active–region magnetic structures.

Deep–learning models can learn complex nonlinear relationships from large observational datasets, but their internal decision process is not directly interpretable. As a result, high classification performance does not automatically imply that the model relies on physically meaningful flare precursors. Interpretability methods are therefore needed to examine whether the model uses relevant spatial structures, temporal information, and magnetic–context variables, and to identify possible failure modes in ambiguous cases. Answering these questions may provide insight into what the model considers important in relation to pre–flare signatures.

In this chapter, we address this issue by applying Layer–wise Relevance Propagation (LRP) to the multimodal forecasting model. The analysis focuses on the full model configuration, MCAF, because it combines all observational channels and physical–context variables considered in the previous chapter. To make relevance propagation more tractable, the original long short–term memory (LSTM) block responsible for temporal modeling is replaced by a simpler one–dimensional convolutional (Conv1d) block, while preserving predictive behavior close to that of the original MCAF model. This modified model is then used as the basis for the interpretability analysis.

The goal of the chapter is to move from forecasting performance to physical interpretation. We examine how relevance is allocated between the image branch and the physical–context branch, how relevance is distributed internally between Φ_{tot} and $h_{c,\text{tot}}$, how the relevance changes across image channels and time steps, and how these patterns differ between true positives (*TP*), false positives (*FP*), true negatives (*TN*), and false negatives (*FN*) predictions. In addition to test–set aggregate statistics, selected case studies are analyzed to compare the spatial organization of relevance maps with the corresponding input observations and post hoc physics–based diagnostic maps.

This chapter is organized as follows. Section 7.2 summarizes the dataset, preprocessing, and modified forecasting model used for the interpretability analysis. Section 7.3 describes the LRP propagation rules and the physics–based diagnostic maps, defines the quantitative attribution metrics, and presents the criteria used to select representative samples for the case–study analysis. Section 7.4 presents the results. Finally, Section 7.5 discusses the main conclusions, methodological limitations, and implications of the analysis.

7.2 Data and Model

This chapter builds on the multimodal flare–forecasting framework introduced in Chapter 6. The input data, dataset construction, labeling strategy, train/validation/test split, and preprocessing are therefore the same as those used in the previous phase. The model used in this chapter also follows the same general architecture as the model introduced in the previous chapter, with some modules modified to adapt it to the goal of the present phase. In the present chapter, however, the purpose is different: rather than evaluating predictive performance, we use the trained forecasting model as the basis for an interpretability analysis. In the following sections, we summarize the most relevant methodological aspects.

7.2.1 Dataset and Preprocessing

The input data consist of LoS magnetograms and continuum intensity images from SDO/HMI, together with extreme–ultraviolet (EUV) images from

SDO/AIA at 193 Å and 304 Å. These observations provide complementary views of active–region structure across different atmospheric layers, from the photosphere to the chromosphere, transition region, and corona. In addition, two scalar parameters from the Space–weather HMI Active Region Patches (SHARP) data product are included: the total unsigned magnetic flux, Φ_{tot}, and the total unsigned current helicity, $h_{c,\mathrm{tot}}$. These quantities were selected as global descriptors of active–region size, magnetic–field strength, and non-potentiality.

The positive class was constructed from the event list compiled by Li et al. (2021), which contains flare events of class C5.0 and above. For each event, a 36 min observational window was extracted, sampled at a 12 min cadence, and ending 2 h before flare onset. At each time step, the corresponding SHARP parameters were retrieved and temporally matched to the nearest HMI and AIA observations. The negative class was constructed from flare–quiet active regions (ARs) by selecting sequences without any flaring activity for at least 3 h before and after the end of the sequence. To reduce projection effects and ensure data quality, only ARs within $\pm 45^{\circ}$ of the central meridian were considered, and sequences containing at least one frame intersecting the solar limb were excluded. After filtering, the final dataset contains 562 flare sequences and 644 non–flare sequences.

The full–disk images were spatially aligned by reprojecting them onto the AIA 304 Å world coordinate system (WCS), ensuring pixel–wise correspondence across image channels. AR cutouts were extracted from the full–disk images using geometric information contained in the SHARP data product about the location and extent of each AR. To extract the AR patches, we used a reference size of 512×512 pixels. Smaller ARs were centered on a frame of this reference size, whereas larger ARs retained their full extent along the corresponding dimension in order to preserve their morphology. Since this procedure produced cutouts of different sizes, random cropping to 512×512 pixels was applied during training, serving both as input–size standardization and as data augmentation. Image channels and scalar parameters were normalized using statistics computed from the training set only. The dataset was split at the AR level into non–overlapping training, validation, and test sets with approximate proportions of 70/15/15%, ensuring that no AR appeared in more than one subset.

7.2.2 Model

The forecasting model used in this chapter follows the hybrid deep–learning architecture introduced in Chapter 6. The model predicts whether an AR will produce a flare of class $\geq$ C5.0 within the next 2 h, using the preceding 36 min of active–region evolution as input. Each time step consists of the image modalities described in Section 6.2.1, optionally combined with the scalar SHARP parameters. This multimodal and multiwavelength formulation allows the model to combine the evolving morphology observed in the image sequence with global magnetic information provided by the physical–parameter time series. As in Chapter 6, three input configurations

are considered: model M, which uses only LoS magnetograms; model MCA, which combines LoS magnetograms, continuum intensity images, and EUV observations; and model MCAF, which further includes the SHARP physical parameters. In the present chapter, however, the interpretability analysis is based solely on model MCAF. For this reason, the following description focuses on the architecture of the full multimodal model.

Each image in the input sequence is individually processed by a convolutional encoder originally trained as part of an autoencoder. The encoder consists of five two–dimensional convolutional blocks that progressively reduce each four–channel 512×512 input frame to a compact latent representation. Global average pooling is then applied to the final feature maps to obtain a fixed–length feature vector for each frame. In parallel, the scalar physical parameters at the three time steps are mapped through a fully connected layer to feature vectors with the same dimensionality as the image–derived features. The image–derived and parameter–derived representations are then fused through element–wise addition, producing a physical–context–aware feature vector at each temporal step, denoted by $\mathbf{C} = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3) \in \mathbb{R}^{3\times512}$.

The temporal evolution of this sequence of fused feature vectors is then processed by two Conv1d blocks. These blocks preserve the temporal dimension while reducing the feature dimension from 512 to 8, producing a representation in $\mathbb{R}^{3\times8}$. The output of the temporal block is then averaged over the temporal dimension to produce a vector in $\mathbb{R}^8$, which is passed to a fully connected classifier with rectified linear units (ReLU) activation, followed by a sigmoid output that returns the predicted probability of flare occurrence.

In the original forecasting model introduced in Chapter 6, temporal evolution was modeled using a two–layer LSTM network. For the interpretability analysis presented in this chapter, this recurrent block was replaced by two Conv1d blocks. This modification was introduced for two main reasons. First, convolutional layers are computationally simpler and contain fewer trainable parameters than the original recurrent layers. Specifically, the original two–layer LSTM, with input size 512 and hidden size 32, contains 78,336 trainable parameters, whereas the two Conv1D layers, with feature dimensions $512 \rightarrow 8 \rightarrow 8$, kernel size 3, and bias terms, contain 12,496 trainable parameters. Second, although relevance–propagation methods have been extended to recurrent architectures, their implementation is more complex than in convolutional and fully connected networks. Replacing the LSTM with a Conv1D block therefore simplifies the implementation of LRP while preserving the ability of the model to combine information across the three time steps.

The modified architecture was trained and evaluated using the same dataset and training methodology as the previous forecasting framework, with the temporal block being the only architectural change. Because the temporal architecture is different, the resulting confusion–matrix counts are not expected to be identical to those reported for the original LSTM–based models in Chapter 6. Relative to the results reported in Table 6.4, Model M produces two fewer false positives and two additional true negatives, while its true–positive and false–negative counts remain unchanged. Model MCA

produces one fewer false positive and one additional true negative, again with unchanged true–positive and false–negative counts. In the full multimodal MCAF configuration, the modified architecture produces one fewer true positive and one additional false negative, while the false–positive and true–negative counts remain unchanged. Thus, the architectural modification leads to small changes in the classification outcomes across the three input configurations. Since the predictive behavior remains comparable, the modified architecture provides a suitable basis for the relevance analysis. The confusion–matrix counts obtained on the test set are summarized in Table 7.1.

Model	TP	FP	TN	FN
Model M	72	10	87	10
Model MCA	77	11	86	5
Model MCAF	79	11	86	3

TABLE 7.1: Confusion–matrix counts for the model configurations used in the interpretability analysis.

7.3 Methodology for Interpretability Analysis

This section introduces the methodology used to interpret the decisions of the forecasting model. Since LRP was already described in Chapter 3, we only specify here the propagation rules adopted for the modified architecture. We then introduce the physics–based diagnostic maps used as post hoc references to assess whether the regions highlighted by the attribution maps are spatially associated with physically meaningful magnetic structures. Next, we define the quantitative attribution metrics used to study relevance allocation. Finally, we describe the criteria used to select representative *TP*, *FP*, *TN*, and *FN* samples for the case–study analysis.

7.3.1 Attribution Method

To interpret the model predictions, we use LRP, following the general formulation introduced in Section 3.7. In brief, LRP redistributes the prediction score backward through the network, assigning relevance values to the input variables according to their contribution to the final output. This procedure produces relevance tensors with the same dimensions as the corresponding inputs: $\mathbf{R}_{\text{img}} \in \mathbb{R}^{3\times4\times512\times512}$ for the image branch, and $\mathbf{R}_{\text{ctx}} \in \mathbb{R}^{3\times2}$ for the physical–context branch. In the present chapter, these attributions are used to identify which regions, time steps, and data modalities contribute most strongly to the flare/no–flare decision.

The specific propagation rules adopted for each layer of the network are summarized in Table 7.2. The selection of these rules reflects the computational role of each layer and the need to preserve numerical stability while obtaining physically interpretable relevance tensors. The α–β rule with $\alpha = 1.5$

and $\beta = 0.5$ is applied to the first convolutional layer in order to emphasize spatially localized positive contributions while still allowing a limited contribution from inhibitory terms. By contrast, deeper convolutional and linear layers are assigned the ϵ–rule with $\epsilon = 10^{-5}$ in order to stabilize relevance redistribution and suppress numerical artifacts in low–activation regions. Operations that mainly preserve or redistribute information, such as averaging layers, nonlinear activations, normalization, and dropout, are handled through routing or pass–through rules. This layer–specific assignment provides a compromise between faithful attribution and stable relevance propagation throughout the network.

LRP Rule	**Applied To**	**Parameters**
α–β rule	First Conv2D layer $(4 \to 64)$	$\alpha = 1.5,\ \beta = 0.5$
ϵ–rule	Conv2D layers $(64 \to 128 \to 256 \to 512 \to 512)$	$\epsilon = 10^{-5}$
Weighted routing	AdaptiveAvgPool2D	–
ϵ–rule	Context mixing layer, Linear $(2 \to 512)$	$\epsilon = 10^{-5}$
ϵ–rule	Temporal encoder, Conv1D $(512 \to 8 \to 8)$	$\epsilon = 10^{-5}$
ϵ–rule	Classifier, Linear $(8 \to 8 \to 1)$	$\epsilon = 10^{-5}$
Uniform routing	Temporal average layer	–
Relevance cut–off	Sigmoid output	–
Pass–through	ReLU, BatchNorm1D, Dropout	–

TABLE 7.2: Layer–specific propagation rules adopted for the modified forecasting model used in this chapter. The rules are assigned according to the role of each layer in order to preserve numerical stability and produce interpretable relevance maps.

The main methodological choices concern the ϵ–rule used for the deeper convolutional and linear layers and the α–β rule used for the input–facing convolutional layer. In the case of the ϵ–rule, changing the value of ϵ modifies the strength of the numerical stabilization and can therefore affect the quantitative relevance values, particularly when the denominator of the propagation rule is close to zero. For sufficiently small values of ϵ, however, the rule approaches the basic z–rule away from near–zero denominators. A more substantive choice concerns the α–β rule, because different admissible parameterizations, subject to the constraint $\alpha - \beta = 1$, correspond to different treatments of excitatory and inhibitory contributions. For example, changing from $(\alpha, \beta) = (1.5, 0.5)$ to $(\alpha, \beta) = (1, 0)$ would change the propagation from one that retains both excitatory and inhibitory contributions to one that suppresses inhibitory contributions entirely.

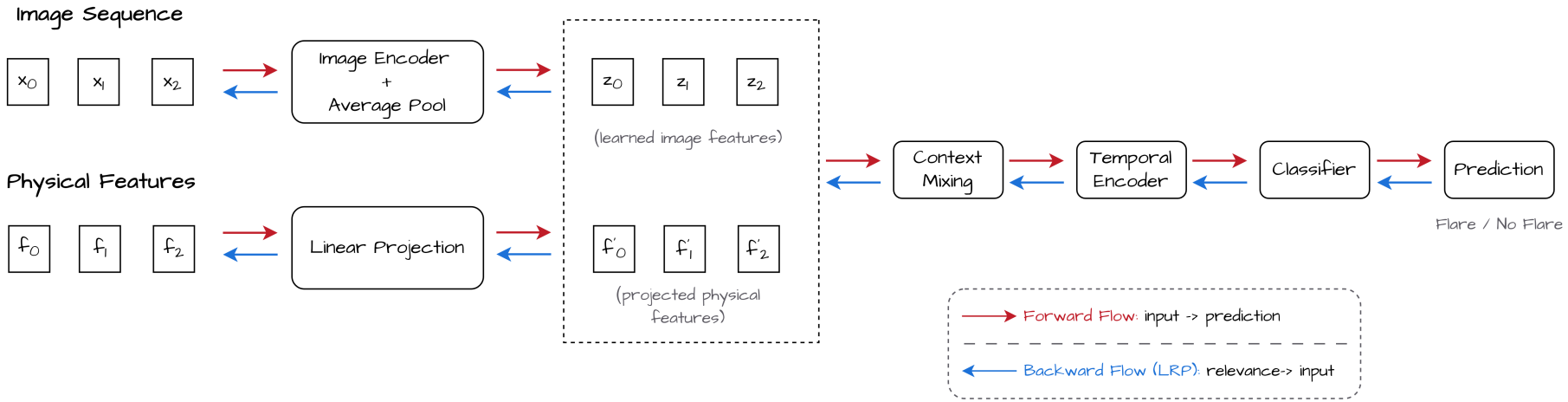


FIGURE 7.1: Schematic illustration of the forward prediction pass and backward relevance propagation used in the interpretability analysis. Red arrows indicate the forward flow from the image sequence and physical–parameter time series to the prediction, while blue arrows indicate the backward LRP flow from the prediction back to the input–level relevance tensors. The dashed box marks the shared latent representation where learned image and physical–parameter features are combined.

A systematic sensitivity analysis with respect to the value of ϵ and to alternative α–β parameterizations was not performed. Consequently, the quantitative relevance allocations reported in this chapter should be interpreted as conditional on the adopted LRP configuration.

Figure 7.1 schematically illustrates the dual flow of information in the model. The red arrows represent the forward pass, in which the image sequence and physical–parameter time series are processed to produce a flare/no–flare prediction. The blue arrows represent the backward relevance propagation, in which the model output is used as the initial relevance and propagated backward toward the inputs. The model has two input branches: an image branch, in which each image in the sequence is processed by the encoder and average–pooling module to obtain learned image features, and a physical–context branch, in which the physical–parameter time series is linearly projected to obtain features with the same dimensionality as the image features. These two representations are fused by the context–mixing module before being processed by the temporal encoder and classifier. During relevance propagation, relevance is redistributed to both branches and then propagated back to the corresponding input–level variables.

7.3.2 Physics–based Diagnostic Maps

To assess whether the regions highlighted by the attribution maps are spatially associated with physically meaningful structures, we compare them with three physics–based diagnostic maps derived from the vector photospheric magnetic field: the free magnetic energy density, ρ_{free}, the vertical current helicity, h_c, and the horizontal gradient of the LoS magnetic field, $|\nabla B_{\text{LoS}}|$.

The free magnetic energy density is defined as

$$\rho_{\text{free}} \propto \left(\mathbf{B}^{\text{obs}} - \mathbf{B}^{\text{pot}}\right)^2, \tag{7.1}$$

where $\mathbf{B}^{\text{obs}}$ and $\mathbf{B}^{\text{pot}}$ are the observed and potential photospheric magnetic fields, respectively. This quantity is used as a proxy for the excess magnetic energy stored in the observed magnetic–field configuration relative to the corresponding potential field, and therefore as an indicator of magnetic nonpotentiality and energy available for release.

The vertical current helicity is defined as

$$h_c = B_z J_z, \tag{7.2}$$

where the vertical current density is given by

$$J_z = \frac{1}{\mu_0}\left(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y}\right). \tag{7.3}$$

This quantity characterizes the coupling between the vertical magnetic field and electric currents, and is therefore interpreted as a measure of magnetic twist and shear in the photospheric magnetic field.

Finally, the horizontal gradient of the LoS magnetic field is defined as

$$|\nabla B_{\mathrm{LoS}}| = \sqrt{\left(\frac{\partial B_{\mathrm{LoS}}}{\partial x}\right)^2 + \left(\frac{\partial B_{\mathrm{LoS}}}{\partial y}\right)^2}, \tag{7.4}$$

which quantifies the strength of spatial magnetic–field variations and highlights compact polarity separations and sharp magnetic transitions.

These quantities were selected because they characterize different physical properties associated with flare productivity: magnetic nonpotentiality, current–carrying and twisted field configurations, and strong small–scale magnetic gradients. Their selection is also motivated by the forecasting model developed in Chapter 6, where the total unsigned magnetic flux, Φ_{tot}, and the total unsigned current helicity, $h_{c,\mathrm{tot}}$, were used as global descriptors of active–region size and nonpotentiality. In the present chapter, h_c provides a local counterpart to the helicity–related information encoded in $h_{c,\mathrm{tot}}$, while ρ_{free} and $|\nabla B_{\mathrm{LoS}}|$ provide complementary spatial diagnostics of magnetic energy storage and field complexity.

These maps are not used as model inputs; instead, they serve as post hoc physical references against which the relevance maps can be qualitatively compared. It is important to note that the physics–based diagnostic maps are computed from vector magnetograms in Cylindrical Equal–Area (CEA) projections, whereas the image inputs to the model are kept in the original charge–coupled device (CCD) coordinates. Therefore, the input images and diagnostic maps do not have a strict pixel–to–pixel correspondence. Nevertheless, they provide useful physical context for interpreting whether the relevance patterns are spatially associated with magnetically complex structures.

7.3.3 Definition of Quantitative Attribution Metrics

We define a set of quantitative attribution metrics that summarize how relevance is distributed over the test set. These metrics are used to compare relevance allocation across classification outcomes, input branches, internally in physical–context variables, temporal steps, image channels.

For a single sample, the image–branch relevance tensor is denoted by

$$\mathbf{R}_{\mathrm{img}} \in \mathbb{R}^{3\times 4\times 512\times 512}, \tag{7.5}$$

where the dimensions correspond to time step, image channel, and the two spatial coordinates. The corresponding physical–context relevance tensor is denoted by

$$\mathbf{R}_{\mathrm{ctx}} \in \mathbb{R}^{3\times 2}, \tag{7.6}$$

where the two physical–context variables are the total unsigned magnetic flux, Φ_{tot}, and the total unsigned current helicity, $h_{c,\mathrm{tot}}$.

The total absolute relevance assigned to the image branch is defined as

$$R_{\mathrm{img}}^{|\cdot|} = \sum_{t,c,x,y} \left|\mathbf{R}_{\mathrm{img}}(t,c,x,y)\right|, \tag{7.7}$$

while the total absolute relevance assigned to the physical–context branch is defined as

$$R_{\mathrm{ctx}}^{|\cdot|} = \sum_{t,k} \left| \mathbf{R}_{\mathrm{ctx}}(t,k) \right| . \tag{7.8}$$

The physical–context relevance is further decomposed into the relevance assigned to each scalar physical parameter:

$$R_{\Phi}^{|\cdot|} = \sum_{t} \left| \mathbf{R}_{\mathrm{ctx}}(t,\Phi_{\mathrm{tot}}) \right| , \tag{7.9}$$

and

$$R_{h_c}^{|\cdot|} = \sum_{t} \left| \mathbf{R}_{\mathrm{ctx}}(t,h_{c,\mathrm{tot}}) \right| . \tag{7.10}$$

Absolute relevance is used for the main allocation metrics because it measures the total magnitude of attribution independently of sign, thereby avoiding the cancellation of positive and negative relevance values in signed sums. To quantify the relative allocation of relevance between the two input branches, we define the image relevance fraction as

$$f_{\mathrm{img}}^{|\cdot|} = \frac{R_{\mathrm{img}}^{|\cdot|}}{R_{\mathrm{img}}^{|\cdot|} + R_{\mathrm{ctx}}^{|\cdot|}}, \tag{7.11}$$

and the physical–context relevance fraction as

$$f_{\mathrm{ctx}}^{|\cdot|} = \frac{R_{\mathrm{ctx}}^{|\cdot|}}{R_{\mathrm{img}}^{|\cdot|} + R_{\mathrm{ctx}}^{|\cdot|}}. \tag{7.12}$$

By construction,

$$f_{\mathrm{img}}^{|\cdot|} + f_{\mathrm{ctx}}^{|\cdot|} = 1. \tag{7.13}$$

The balance between the two branches is summarized by

$$\Delta_{\mathrm{img-ctx}}^{|\cdot|} = f_{\mathrm{img}}^{|\cdot|} - f_{\mathrm{ctx}}^{|\cdot|}. \tag{7.14}$$

Positive values of $\Delta_{\mathrm{img-ctx}}^{|\cdot|}$ indicate image–dominant attribution, whereas negative values indicate physical–context–dominant attribution.

To examine the internal allocation of relevance within the physical–context branch, we define

$$f_{\Phi|\mathrm{ctx}}^{|\cdot|} = \frac{R_{\Phi}^{|\cdot|}}{R_{\mathrm{ctx}}^{|\cdot|}}, \tag{7.15}$$

and

$$f^{|\cdot|}_{h_c|\mathrm{ctx}} = \frac{R^{|\cdot|}_{h_c}}{R^{|\cdot|}_{\mathrm{ctx}}}. \tag{7.16}$$

Since the physical–context branch contains only Φ_{tot} and $h_{c,\mathrm{tot}}$, these fractions satisfy

$$f^{|\cdot|}_{\Phi|\mathrm{ctx}} + f^{|\cdot|}_{h_c|\mathrm{ctx}} = 1, \tag{7.17}$$

provided that $R^{|\cdot|}_{\mathrm{ctx}} \neq 0$.

In addition to absolute relevance, we also quantify the sign of the relevance contributions. For a generic branch b, the positive and negative relevance magnitudes are defined as

$$R^{+}_{b} = \sum_{j \in b} \max(R_j, 0), \tag{7.18}$$

and

$$R^{-}_{b} = \sum_{j \in b} \max(-R_j, 0), \tag{7.19}$$

where j indexes the elements of the corresponding relevance tensor. The positive and negative relevance fractions are then defined as

$$f^{+}_{b} = \frac{R^{+}_{b}}{R^{+}_{b} + R^{-}_{b}}, \tag{7.20}$$

and

$$f^{-}_{b} = \frac{R^{-}_{b}}{R^{+}_{b} + R^{-}_{b}}. \tag{7.21}$$

These quantities describe whether the relevance assigned to a branch is mainly supportive or inhibitory with respect to the model output.

To analyze the temporal distribution of relevance, we compute branch–level relevance separately at each time step. For the image branch,

$$R^{|\cdot|}_{\mathrm{img}}(t) = \sum_{c,x,y} \left| \mathbf{R}_{\mathrm{img}}(t, c, x, y) \right|, \tag{7.22}$$

whereas for the physical–context branch,

$$R^{|\cdot|}_{\mathrm{ctx}}(t) = \sum_{k} \left| \mathbf{R}_{\mathrm{ctx}}(t, k) \right|. \tag{7.23}$$

These quantities are computed for the three input time steps t_0, t_1, and t_2.

Finally, to compare the image channels, we define the total absolute relevance assigned to channel c as

$$R^{|\cdot|}_{c} = \sum_{t,x,y} \left| \mathbf{R}_{\mathrm{img}}(t, c, x, y) \right|. \tag{7.24}$$

The four channels correspond to HMI LoS magnetogram (B_{LoS}), HMI continuum intensity (I_c), AIA 193 Å, and AIA 304 Å.

After these metrics are computed for each test sample, they are grouped according to the classification outcome: *TP*, *FP*, *TN*, and *FN*. The distributions are summarized using the median, first quartile (Q_1), third quartile (Q_1), and interquartile range ($IQR = Q_3 - Q_1$). These quantities provide the quantitative basis for the analyses presented in the following sections.

7.3.4 Selection of Samples for Interpretation

The case–study analysis was based on four representative test–set samples, corresponding to one example from each classification outcome. The objective was not to provide an exhaustive visual inspection of all test samples, but to select representative cases that could illustrate how the attribution patterns differ between correct classifications and misclassifications.

The samples were selected by comparing the predictions of the three forecasting models: model M, which uses only LoS magnetograms; model MCA, which combines LoS magnetograms, continuum intensity images, and EUV observations; and model MCAF, which further includes the SHARP physical parameters. For the correctly classified examples, the true positive and true negative samples were selected among cases that were classified correctly by all three models and for which the predicted probabilities indicated high confidence. For the incorrectly classified example, the false positive and false negative samples were selected among cases that were misclassified by all three models. This criterion was adopted to focus the visual interpretation on robust model behavior, rather than on isolated errors associated with only one model configuration.

Table 7.3 reports the test-set samples misclassified by all three models, which were used as the candidate set for selecting the false positive and false negative examples discussed in the case-study analysis.

The final set of samples selected for visual interpretation is summarized in Table 7.4. The true positive case, HARP 5298, was selected as a confidently detected flaring region with high predicted flare probability across all three models. The true negative case, HARP 2779, was selected as a confidently detected quiet region with low predicted flare probability across all three models. The false positive case, HARP 3730, was selected from the consensus misclassifications set, for which all three models assigned flare probabilities above the decision threshold despite the absence of a flare in the prediction window. The false negative case, HARP 2809, was selected from the same set, for which all three models assigned low flare probabilities despite the occurrence of an M–class flare.

These four samples are used in Section 7.4.3 to compare the LRP relevance maps with the corresponding input observations and post hoc physics–based diagnostic maps. This selection strategy ensures that the case studies are tied to consistent behavior across the model configurations, while still covering both correct and incorrect classifications.

Outcome	HARP	Sequence Start	Sequence End	p_M	p_{MCA}	p_{MCAF}
FN	1603	2012.04.27 06:00:00	2012.04.27 06:24:00	0.151	0.357	0.351
FN	2809	2013.05.31 17:36:00	2013.05.31 18:00:00	0.101	0.143	0.029
FP	2605	2013.04.07 00:24:00	2013.04.07 00:48:00	0.573	0.627	0.648
FP	3730	2014.02.14 23:00:00	2014.02.14 23:24:00	0.587	0.829	0.625
FP	3793	2014.02.27 23:48:00	2014.02.28 00:12:00	0.505	0.819	0.885
FP	6558	2016.05.20 21:36:00	2016.05.20 22:00:00	0.655	0.760	0.778

TABLE 7.3: Consensus misclassifications in the test set. The table includes only samples misclassified by all three model configurations. The quantities p_M, p_{MCA}, and p_{MCAF} denote the predicted flare probabilities produced by model M (LoS magnetogram–only), model MCA (LoS magnetogram + continuum + EUV), and model MCAF (LoS magnetogram + continuum + EUV + physical parameters), respectively.

Case	HARP	Sequence Start	Sequence End	p_M	p_{MCA}	p_{MCAF}
TP	5298	2015.03.12 11:48:00	2015.03.12 12:12:00	0.853	0.977	0.989
TN	2779	2013.06.03 09:24:00	2013.06.03 09:48:00	0.061	0.090	0.014
FP	3730	2014.02.14 23:00:00	2014.02.14 23:24:00	0.587	0.829	0.625
FN	2809	2013.05.31 17:36:00	2013.05.31 18:00:00	0.101	0.143	0.029

TABLE 7.4: Representative samples selected for the case–study interpretation. The selected examples include one *TP*, one *TN*, one *FP*, and one *FN* case. The reported values are the predicted flare probabilities from the three model configurations.

7.4 Results

This section presents the interpretability analysis of the modified multimodal forecasting model. The goal is not to re–evaluate the predictive performance of the model, which was already discussed in Chapter 6, but to examine how the model forms its decisions. To that end, we first analyze the global relevance statistics for the test set in order to understand how the model distributes relevance across input modalities and whether the physical–context branch is biased toward one of the two global physical descriptors. We then qualitatively compare the relevance patterns of selected case studies representing the four classification outcomes.

7.4.1 Quantitative Attribution Analysis

Figure 7.2 summarizes the allocation of absolute relevance across the test set, grouped by classification outcome, for the MCAF model. In contrast to an analysis based on total absolute relevance, the quantities shown here describe how the available absolute relevance is fractionally distributed across the different input modalities and among the physical–context variables. This distinction is important because the relevance tensors are obtained by propagating the scalar model output backward through the network. Since the model output is a probability–like value between 0 and 1, samples classified as negative tend to be associated with output values close to 0, whereas samples classified as positive tend to be associated with output values close to 1. As a consequence, samples classified as negative generally have smaller total absolute relevance values than samples classified as positive. Directly comparing total absolute relevance across classification outcomes would therefore partly reflect the output and confidence of the model, rather than only the relevance of different input modalities.

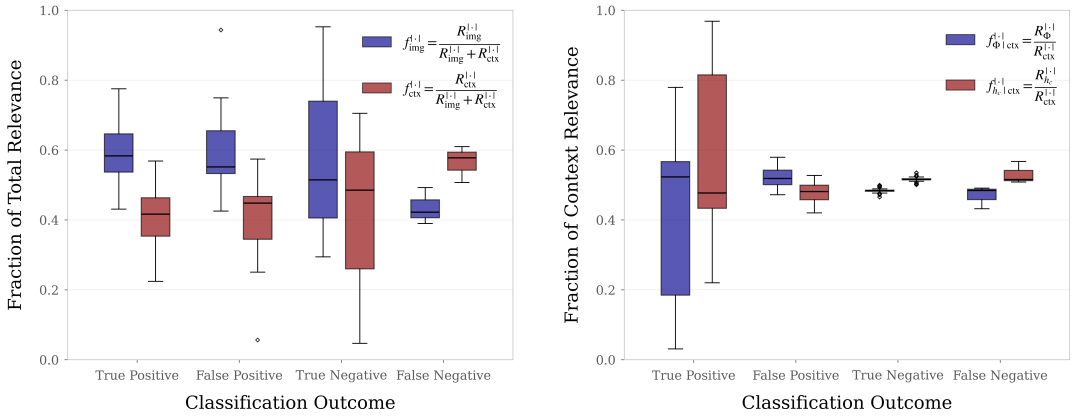


FIGURE 7.2: Distribution of fractional relevance–allocation metrics grouped by classification outcome for the MCAF model. The left panel compares the fraction of total absolute relevance assigned to the image and physical–context branches. The right panel compares the internal fraction of physical–context relevance assigned to Φ_{tot} and $h_{c,\text{tot}}$.

For this reason, the analysis focuses on relevance allocation, defined as the fraction of total absolute relevance assigned to each input modality (see Section 7.3.3). The left panel of Figure 7.2 compares the image relevance fraction, $f^{|\cdot|}_{\text{img}}$, with the physical–context relevance fraction, $f^{|\cdot|}_{\text{ctx}}$. The right panel analyzes the physical–context branch internally by comparing the magnetic–flux relevance fraction, $f^{|\cdot|}_{\Phi|\text{ctx}}$, with the current–helicity relevance fraction, $f^{|\cdot|}_{h_c|\text{ctx}}$. The corresponding summary statistics are reported in Table 7.5.

The first result is that correctly detected flaring samples are predominantly image–driven. For *TP*, the median image relevance fraction is 0.583, compared with 0.417 for the physical–context relevance fraction. This corresponds to a positive median image–context difference of 0.146. Therefore, when the model correctly identifies a flaring AR, the decision is generally supported more strongly by the spatio–temporal image representation than by the physical–parameter time series.

TN show a more balanced relevance allocation. Their median image relevance fraction is 0.515, while the median physical–context relevance fraction is 0.485, giving a median image–context difference of 0.026. In contrast to the *TP* case, where the interquartile ranges of the image and physical–context relevance fractions do not overlap, the corresponding *TN* intervals overlap substantially. For *TN*, the image relevance fraction has an interquartile range of 0.405–0.740, while the physical–context relevance fraction has an interquartile range of 0.260–0.595, producing an overlap interval of 0.405–0.595. This overlap indicates that, among correctly classified quiet samples, the relative contribution of the two modalities is less clearly separated.

In some correctly classified quiet samples, the relevance allocation is dominated by the image sequences, whereas in others the physical–parameter time series receives the larger relative contribution. This suggests that quiet–region classifications are not explained by a single, uniform relevance–allocation pattern. Rather, the model may reject flare productivity using different combinations of spatio–temporal image information and scalar magnetic descriptors, depending on the structure of the AR.

The second result concerns the internal distribution of relevance within the physical–context branch. For *TP*, the median magnetic–flux relevance fraction is $f^{|\cdot|}_{\Phi|\text{ctx}} = 0.523$, compared with $f^{|\cdot|}_{h_c|\text{ctx}} = 0.477$ for the current–helicity relevance fraction. This indicates a slight median dominance of Φ_{tot} within the physical–context branch for correctly detected flares. However, this result should be interpreted together with the modality–level allocation discussed above: in *TP*, the physical–context branch receives a smaller fraction of the total absolute relevance than the image branch. Thus, although Φ_{tot} contributes more than $h_{c,\text{tot}}$ within the physical–context branch, the dominant support for the positive classification remains image–based.

Metric	TP	FP	TN	FN
N	79	11	86	3
$f_{\text{img}}^{\lvert\cdot\rvert}$	0.583 [0.537, 0.646]	0.552 [0.533, 0.655]	0.515 [0.405, 0.740]	0.422 [0.406, 0.458]
$f_{\text{ctx}}^{\lvert\cdot\rvert}$	0.417 [0.354, 0.463]	0.448 [0.345, 0.467]	0.485 [0.260, 0.595]	0.578 [0.542, 0.594]
$\Delta_{\text{img}-\text{ctx}}^{\lvert\cdot\rvert}$	0.146 [0.064, 0.262]	0.090 [0.057, 0.286]	0.026 [-0.166, 0.454]	-0.136 [-0.165, -0.074]
$f_{\Phi\lvert\text{ctx}}^{\lvert\cdot\rvert}$	0.523 [0.185, 0.567]	0.519 [0.501, 0.542]	0.484 [0.482, 0.485]	0.484 [0.458, 0.488]
$f_{h_c\lvert\text{ctx}}^{\lvert\cdot\rvert}$	0.477 [0.433, 0.815]	0.481 [0.458, 0.499]	0.516 [0.515, 0.518]	0.516 [0.512, 0.542]

TABLE 7.5: Summary statistics of the relevance–allocation metrics grouped by classification outcome for the MCAF model. Values are reported as median $[Q_1, Q_3]$, except for N, which gives the number of samples in each group. The metric $\Delta_{\text{img}-\text{ctx}}^{\lvert\cdot\rvert}$ is defined as the base–10 logarithmic difference between the image and physical-context relevance fractions.

For *TN*, the internal physical–context allocation shows the opposite tendency. The median current–helicity relevance fraction is $f^{|\cdot|}_{h_c|\mathrm{ctx}} = 0.516$, compared with $f^{|\cdot|}_{\Phi|\mathrm{ctx}} = 0.484$ for the magnetic–flux relevance fraction. In addition, these distributions are much narrower than in the *TP* case, with interquartile ranges of 0.482–0.485 and 0.515–0.518 for $f^{|\cdot|}_{h_c|\mathrm{ctx}}$ and $f^{|\cdot|}_{\Phi|\mathrm{ctx}}$, respectively. This indicates that, when the physical–parameter time series contributes more to correctly classified quiet samples, its relevance is consistently weighted slightly more toward current helicity than toward magnetic flux.

The *FP* and *FN* groups are much smaller, with $N = 11$ and $N = 3$, respectively, and should therefore be interpreted only descriptively. *FP* show an image–dominant tendency, with a median image relevance fraction of 0.552, a median physical–context relevance fraction of 0.448, and $\Delta^{|\cdot|}_{\mathrm{img-ctx}} = 0.090$. By contrast, *FN* show a context–dominant tendency, with a median physical–context relevance fraction of 0.578, a median image relevance fraction of 0.422, and $\Delta^{|\cdot|}_{\mathrm{img-ctx}} = -0.136$. Although these trends are consistent with the idea that *FP* may be driven by image patterns interpreted as flare–relevant, whereas missed flares may rely more on the physical–parameter time series in the absence of a sufficiently strong support from the image representation, the small number of samples prevents drawing a statistically robust conclusion.

Tables 7.6–7.8 further decompose the total absolute relevance across the image channels and time steps for the full test set (see Section 7.3.3). Across the image channels, the total absolute relevance is largest for the EUV channels, especially AIA 304 Å, which has the highest median relevance (1.540×10^{-3}), followed by AIA 193 Å (8.962×10^{-4}), I_c (4.149×10^{-4}), and B_{LoS} (4.176×10^{-5}). The temporal decomposition shows that the relevance is not uniformly distributed across time. For the image sequences, the largest median relevance occurs at t_0 (1.143×10^{-3}), followed by t_2 (1.065×10^{-3}) and t_1 (7.356×10^{-4}). In contrast, for the physical–parameter time series, the largest median relevance occurs at t_1 (1.022×10^{-3}), followed by t_0 (7.770×10^{-4}) and t_2 (6.101×10^{-4}). This indicates that the model does not use the temporal information in the same way for the two modalities: the image branch places relatively stronger emphasis on the first and last frames of the sequence, whereas the physical–context branch assigns the largest relevance to the central time step.

Statistic	B_{LoS}	I_c	AIA 193	AIA 304
Mean	8.720×10^{-5}	6.747×10^{-4}	1.092×10^{-3}	1.892×10^{-3}
Median	4.176×10^{-5}	4.149×10^{-4}	8.962×10^{-4}	1.540×10^{-3}
Q_1	7.479×10^{-7}	2.098×10^{-5}	4.991×10^{-5}	7.113×10^{-5}
Q_3	1.227×10^{-4}	8.707×10^{-4}	1.425×10^{-3}	2.690×10^{-3}
IQR	1.220×10^{-4}	8.497×10^{-4}	1.375×10^{-3}	2.619×10^{-3}

TABLE 7.6: Summary statistics of the total absolute relevance assigned to each image channel by the MCAF model over the full test set.

Statistic	$R^{\lvert\cdot\rvert}_{\mathrm{img}}(t_0)$	$R^{\lvert\cdot\rvert}_{\mathrm{img}}(t_1)$	$R^{\lvert\cdot\rvert}_{\mathrm{img}}(t_2)$
Mean	1.441×10^{-3}	1.026×10^{-3}	1.280×10^{-3}
Median	1.143×10^{-3}	7.356×10^{-4}	1.065×10^{-3}
Q_1	5.290×10^{-5}	3.809×10^{-5}	5.163×10^{-5}
Q_3	2.000×10^{-3}	1.337×10^{-3}	1.799×10^{-3}
IQR	1.947×10^{-3}	1.299×10^{-3}	1.747×10^{-3}

TABLE 7.7: Summary statistics of the total absolute relevance assigned by the MCAF model to the image sequence at each time step over the full test set.

Statistic	$R^{\lvert\cdot\rvert}_{\mathrm{ctx}}(t_0)$	$R^{\lvert\cdot\rvert}_{\mathrm{ctx}}(t_1)$	$R^{\lvert\cdot\rvert}_{\mathrm{ctx}}(t_2)$
Mean	6.126×10^{-4}	8.273×10^{-4}	5.311×10^{-4}
Median	7.770×10^{-4}	1.022×10^{-3}	6.101×10^{-4}
Q_1	3.876×10^{-5}	5.097×10^{-5}	3.193×10^{-5}
Q_3	1.135×10^{-3}	1.544×10^{-3}	9.280×10^{-4}
IQR	1.096×10^{-3}	1.493×10^{-3}	8.961×10^{-4}

TABLE 7.8: Summary statistics of the total absolute relevance assigned by the MCAF model to the physical–parameter time series at each time step over the full test set.

Metric	TP	FP	TN	FN
N	79	11	86	3
f^+_{img}	0.949 [0.928, 0.958]	0.940 [0.922, 0.960]	0.950 [0.907, 0.969]	0.928 [0.927, 0.942]
f^-_{img}	0.051 [0.042, 0.072]	0.060 [0.040, 0.078]	0.050 [0.031, 0.093]	0.072 [0.058, 0.073]
f^+_{ctx}	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
f^-_{ctx}	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]

TABLE 7.9: Summary statistics of the positive and negative relevance–fraction metrics grouped by classification outcome for the MCAF model. Values are reported as median $[Q_1, Q_3]$, except for N, which gives the number of samples in each group. The metrics f^+ and f^- quantify the fractions of total relevance magnitude associated with positive and negative relevance contributions, respectively.

Because the relevance fractions are relative quantities, the two branches compete for the same total relevance budget. Therefore, the opposite behavior observed between the image branch and the physical–context branch is expected by construction. A more informative point of discussion is where the temporal maximum occurs within each modality. For the image branch, the observed tendency shows higher relevance at t_0 and t_2, corresponding to the initial and final states of the input sequence. This pattern may suggest that the boundary frames provide the most relevant information about short–term changes in the AR structure. In this interpretation, the intermediate frame contributes less unique information because the model can rely more strongly on the contrast, or difference, between the beginning and the end of the sequence when forming its decision. By contrast, for the physical–parameter time series, the central time step receives the largest relevance. This may indicate that the central time step acts as a representative state, similar to a local average over the short–term evolution, especially if the physical parameters vary smoothly across the sequence.

Table 7.9 reports the distribution of positive and negative relevance fractions for the image and physical–context branches. Across all classification outcomes, the image branch is dominated by positive relevance, with median values of f^{+}_{img} between 0.928 and 0.950. The corresponding negative relevance fractions remain small, with median values between 0.050 and 0.072. The physical–context branch shows an even more one–sided behavior, with $f^{+}_{\mathrm{ctx}} = 1.000$ and $f^{-}_{\mathrm{ctx}} = 0.000$ for all outcome groups. This indicates that, in the physical–context branch, the propagated relevance is entirely positive in these aggregated statistics, whereas the image branch retains a small inhibitory component.

7.4.2 Physical–Parameter Distributions by Classification Outcome

To complement the relevance–allocation analysis, Figure 7.3 shows the distributions of the time–aggregated values of the two scalar physical parameters used as physical contextual inputs grouped by classification outcome. In this figure, both quantities are shown after standardization, rather than in their original physical units. For each parameter, the test–set mean is subtracted and the result is divided by the corresponding test–set standard deviation:

$$z_{\Phi_{\mathrm{tot}}} = \frac{\Phi_{\mathrm{tot}} - \mu_{\Phi_{\mathrm{tot}}}}{\sigma_{\Phi_{\mathrm{tot}}}}, \qquad z_{h_{c,\mathrm{tot}}} = \frac{h_{c,\mathrm{tot}} - \mu_{h_{c,\mathrm{tot}}}}{\sigma_{h_{c,\mathrm{tot}}}}. \tag{7.25}$$

This standardization places the two physical parameters on a common scale and makes their distributions directly comparable in terms of deviations from the test–set mean. The dashed horizontal line in Figure 7.3 therefore corresponds to the standardized mean value, $z = 0$.

The standardized distributions show that *TP* are associated with the largest values of both physical parameters. For *TP*, the distribution of $z_{\Phi_{\mathrm{tot}}}$ has a

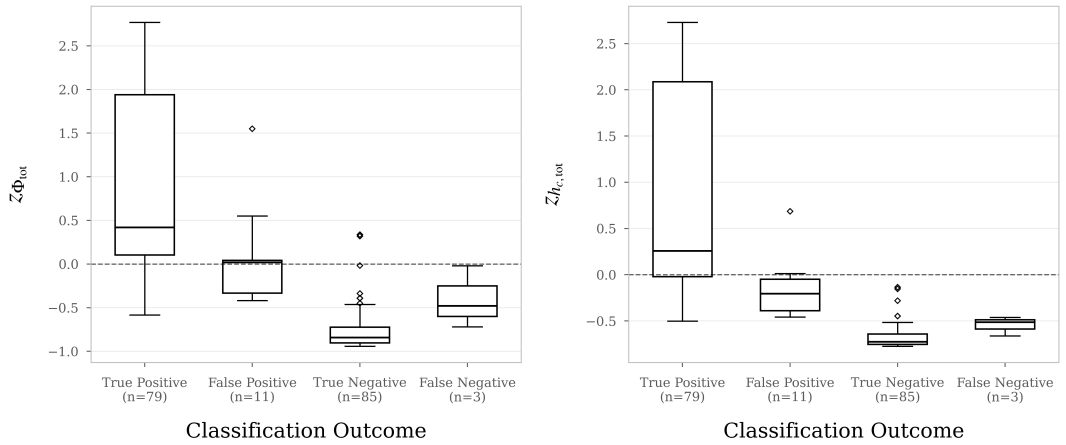


FIGURE 7.3: Distribution of the standardized total unsigned magnetic flux, $z_{\Phi_{tot}}$, and total unsigned current helicity, $z_{h_{c,tot}}$, grouped by classification outcome. The dashed horizontal line indicates the test–set mean after standardization.

positive median and extends to values more than two standard deviations above the test–set mean. A similar behavior is observed for $z_{h_{c,tot}}$, whose *TP* distribution also spans the highest range among the four classification outcomes. This indicates that correctly detected flare–producing ARs are generally characterized by above–average total unsigned magnetic flux and total unsigned current helicity, consistent with their larger magnetic size and stronger nonpotential magnetic configuration. *TN* show the opposite behavior. Their distributions are concentrated below the standardized mean for both parameters. This indicates that correctly classified quiet ARs are generally associated with below–average total unsigned magnetic flux and total unsigned current helicity.

The misclassified groups occupy intermediate regimes, although they must be interpreted cautiously because of their small sample sizes. *FP* show values closer to the standardized mean than *TP*, but they still include cases with positive standardized total unsigned magnetic flux and total unsigned current helicity. This suggests that some false alarms occur in ARs that are not physically trivial, but instead have magnetic properties that partially resemble those of flare–producing ARs. *FN*, in contrast, remain below the *TP* distribution but above *TN* distribution for both standardized parameters. Their values suggest that missed flares may occur in regions whose global total unsigned magnetic flux and total unsigned current helicity are less pronounced, even if they still produce a flare within the prediction window.

7.4.3 Case Studies

The previous subsections analyzed how the MCAF model distributes relevance across input modalities and examined the distributions of the physical parameters associated with each classification outcome. These analyses were performed at the test–set level, with the aim of assessing the overall behavior of the model. We now examine four representative ARs in more detail,

corresponding to one *TP*, one *FP*, one *TN*, and one *FN* case. These examples were selected using the procedure described in Section 7.3.4. The purpose of this analysis is not to establish a pixel–by–pixel correspondence between the relevance maps and the physics–based diagnostic maps. Rather, the goal is to interpret the model behavior qualitatively in terms of AR morphology, relevance distribution, and physical plausibility.

For each case, Figures 7.6–7.9 show the LRP relevance maps over the three input time steps and four image channels: HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å. Positive relevance indicates input regions that support the model output, whereas negative relevance indicates regions that suppress it. Since the propagated output value differs between positive and negative classifications, the case–study discussion focuses primarily on the spatial structure of the relevance patterns, such as compactness, diffuseness, and association with flare–related AR features, rather than on direct comparisons of absolute relevance magnitude across outcomes. Figure 7.4 shows the corresponding final input frame for the same four cases, while Figure 7.5 shows post hoc physics–based diagnostic maps derived from the vector magnetic field. These diagnostic maps provide physical context for the interpretation but are not used as model inputs.

True Positive: HARP 5298

The *TP* case corresponds to HARP 5298, from 2015.03.12 11:48 to 12:12, and is associated with an M3.2 flare. This sample was classified as flaring by all three models, with predicted flare probabilities of $p_M = 0.853$, $p_{MCA} = 0.977$, and $p_{MCAF} = 0.989$. The relevance maps in Figure 7.6 show structured positive relevance across all image channels and time steps. The strongest and most spatially coherent relevance appears in the AIA 304 Å channel, where it follows bright transition–region structures around the AR core. The AIA 193 Å relevance is weaker but still follows extended coronal structures, especially around the central part of the AR. The HMI I_c relevance is concentrated near compact sunspot structures. The HMI B_{LoS} relevance is weaker but still localized around the core concentration of magnetic flux and, notably, appears spatially coherent with the polarity–inversion–line–like structure visible in the final input frame of the sequence shown in Figure 7.4.

The corresponding last–time–step input images in Figure 7.4 and physics–based diagnostic maps in Figure 7.5 indicate that this is a magnetically complex AR. The final time step of the image sequence shows a compact sunspot group, strong magnetic concentrations, and bright coronal and transition–region emission, which are also the structures receiving higher relevance in Figure 7.6. The physics–based maps show enhanced free magnetic energy density, coherent current–helicity structures, and strong magnetic–field gradients concentrated around the AR core. Although these maps do not correspond pixel–by–pixel to the relevance maps, they qualitatively support the broader conclusion that the model assigns relevance to an AR with clear physical signatures commonly associated with flare productivity. This case

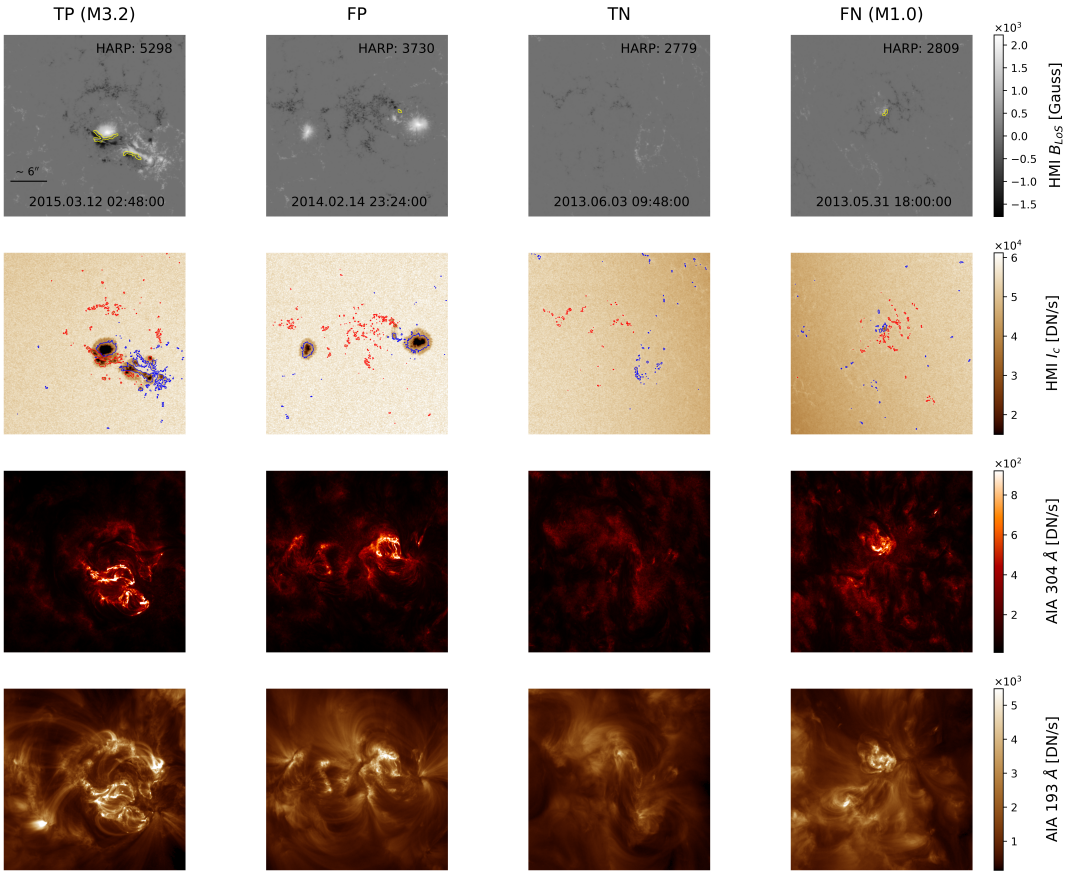


FIGURE 7.4: Final input frame for the four representative ARs used in the case–study analysis. Columns correspond to the *TP*, *FP*, *TN*, and *FN* cases. Rows show HMI B_{LoS}, HMI I_c, AIA 304 Å, and AIA 193 Å. In the B_{LoS} row, the yellow contours indicate the polarity inversion lines (PILs). In the I_c row, red and blue contours indicate positive and negative magnetic polarity, respectively.

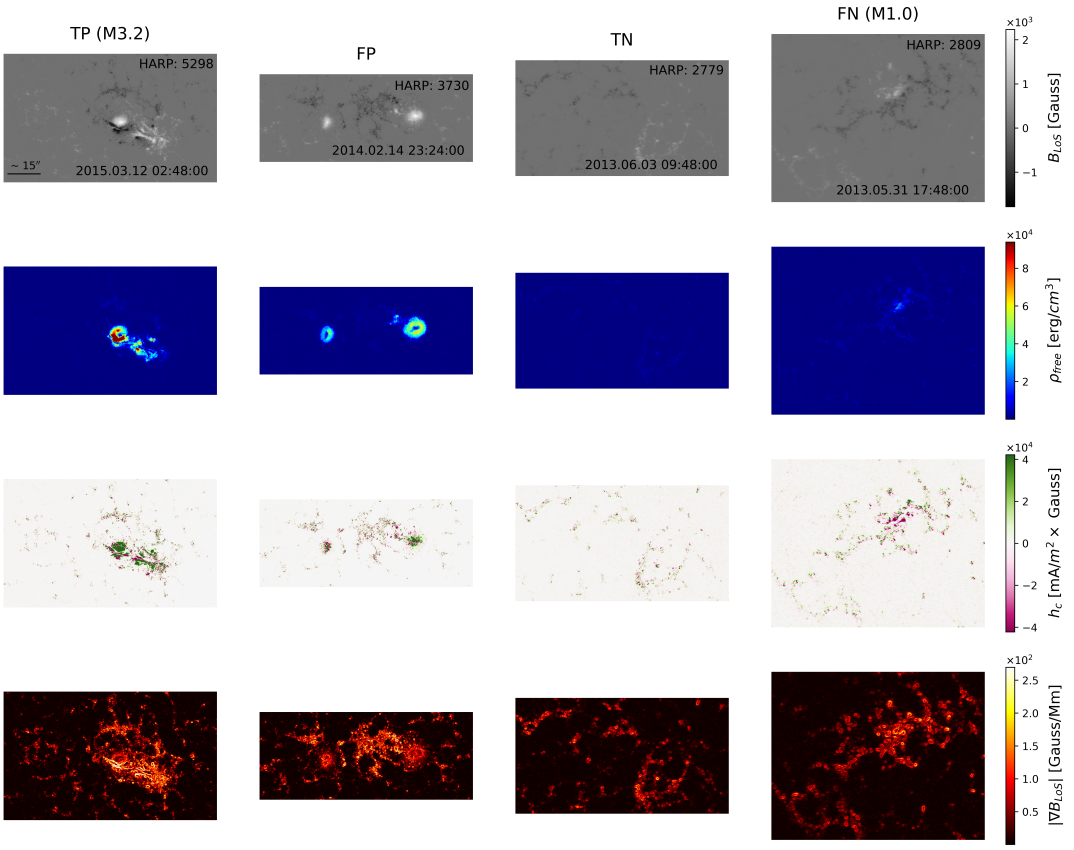


FIGURE 7.5: Physics–based diagnostic maps for the four representative ARs used in the case–study analysis. Columns correspond to the *TP*, *FP*, *TN*, and *FN* cases. Rows show the LoS magnetic field, the free magnetic energy density ρ_{free}, the vertical current helicity h_c, and the horizontal gradient magnitude $|\nabla B_{\text{LoS}}|$.

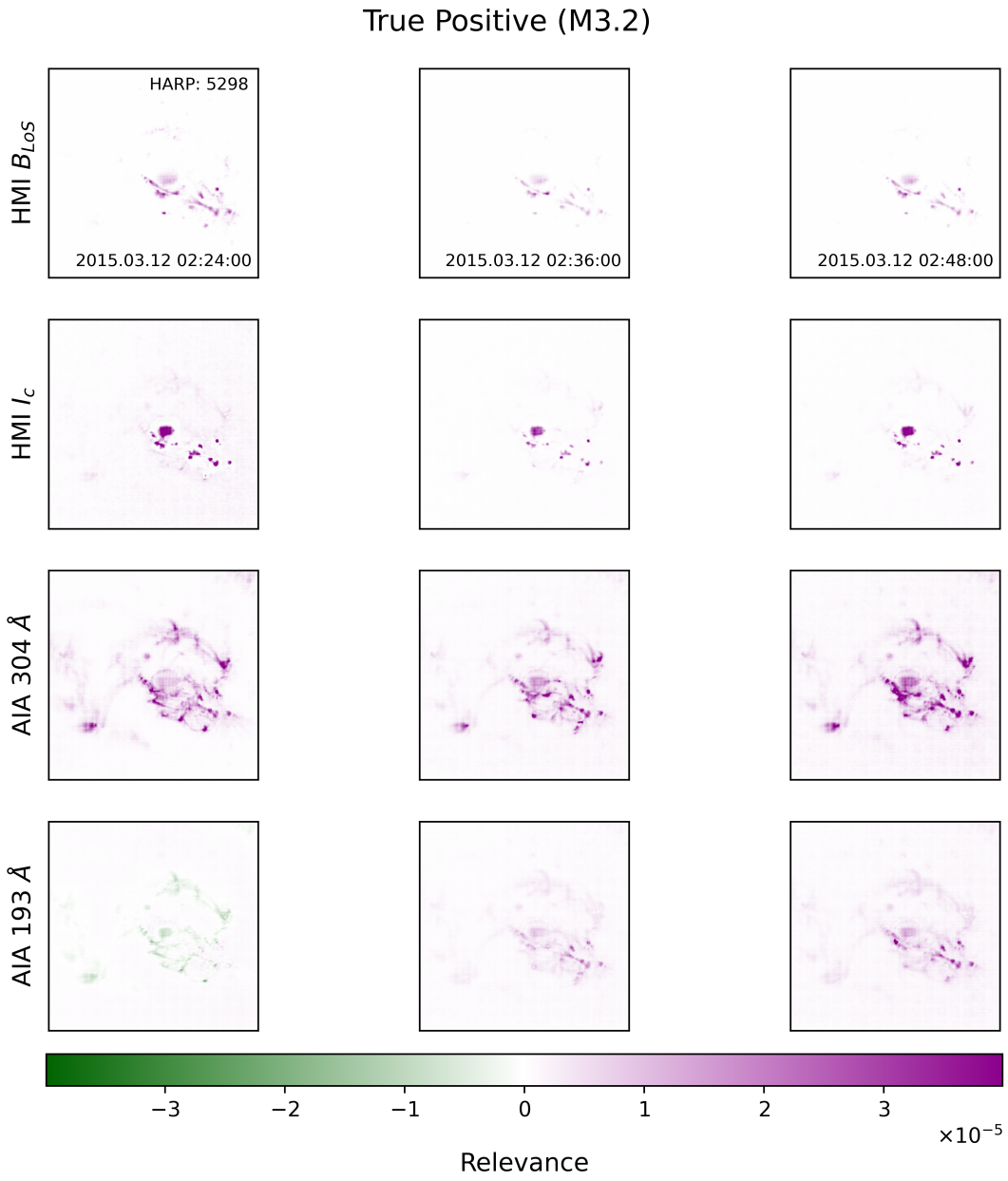


FIGURE 7.6: LRP relevance maps for the *TP* case, HARP 5298, associated with an M3.2 flare. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å.

therefore illustrates a successful prediction in which strong image–based relevance and high physical complexity are consistent with the classification outcome.

False Positive: HARP 3730

The *FP* case corresponds to HARP 3730, from 2014.02.14 23:00 to 23:24, and is a non–flaring AR that was classified as flare–producing by all three models, with predicted flare probabilities of $p_M = 0.587$, $p_{MCA} = 0.829$, and $p_{MCAF} = 0.625$. The relevance maps in Figure 7.7 show non–negligible positive relevance in both photospheric and coronal channels. However, compared with the *TP* case, the relevance pattern is more spatially diffuse: compact regions of high positive relevance are present, but they are embedded in a weak positive background that extends over much of the AR. The HMI I_c relevance is concentrated around two compact sunspot–like structures, while the HMI B_{LoS} relevance is localized around stronger magnetic–flux concentrations, with particular emphasis on a small region that visually corresponds to a polarity–inversion–line–like structure visible in the final input frame of the sequence shown in Figure 7.4. The AIA 304 Å channel shows extended positive relevance over bright transition–region emission, whereas the AIA 193 Å channel shows very diffuse relevance over coronal structures. These patterns indicate that the model responded to spatially organized and visually active structures, but with a relevance distribution less compact than that observed for the correctly detected flare.

The physics–based diagnostic maps in Figure 7.5 show that the free magnetic energy density is localized around compact magnetic structures, but with lower values than those observed for HARP 5298. The current–helicity map shows fragmented positive and negative patches rather than a coherent helicity pattern, and the magnetic–gradient map indicates localized complexity, but without the same degree of organization seen in the *TP* case. This suggests that the model may have responded to active–looking morphology, such as compact photospheric features and bright transition–region emission, even though the AR did not exhibit magnetic–field strengths and magnetic stress comparable to those of HARP 5298. We note that HARP 3730, corresponding to NOAA 11976, later produced a series of small flares ($\leq$ C5.0) between 2014–02–17 and 2014–02–19, as well as an M3.0 flare when the AR was closer to the solar limb.

True Negative: HARP 2779

The *TN* case corresponds to HARP 2779, from 2013.06.03 09:24 to 09:48, and is a non–flaring AR that was correctly classified as quiet by all three models, with predicted flare probabilities of $p_M = 0.061$, $p_{MCA} = 0.090$, and $p_{MCAF} = 0.014$. The relevance maps in Figure 7.8 show a spatially diffuse pattern across the image channels. The HMI B_{LoS} and I_c relevance are weak compared with the AIA relevance and are broadly distributed over low–contrast background structures, consistent with the absence of clear photospheric structures as observed in the final input frame of the sequence

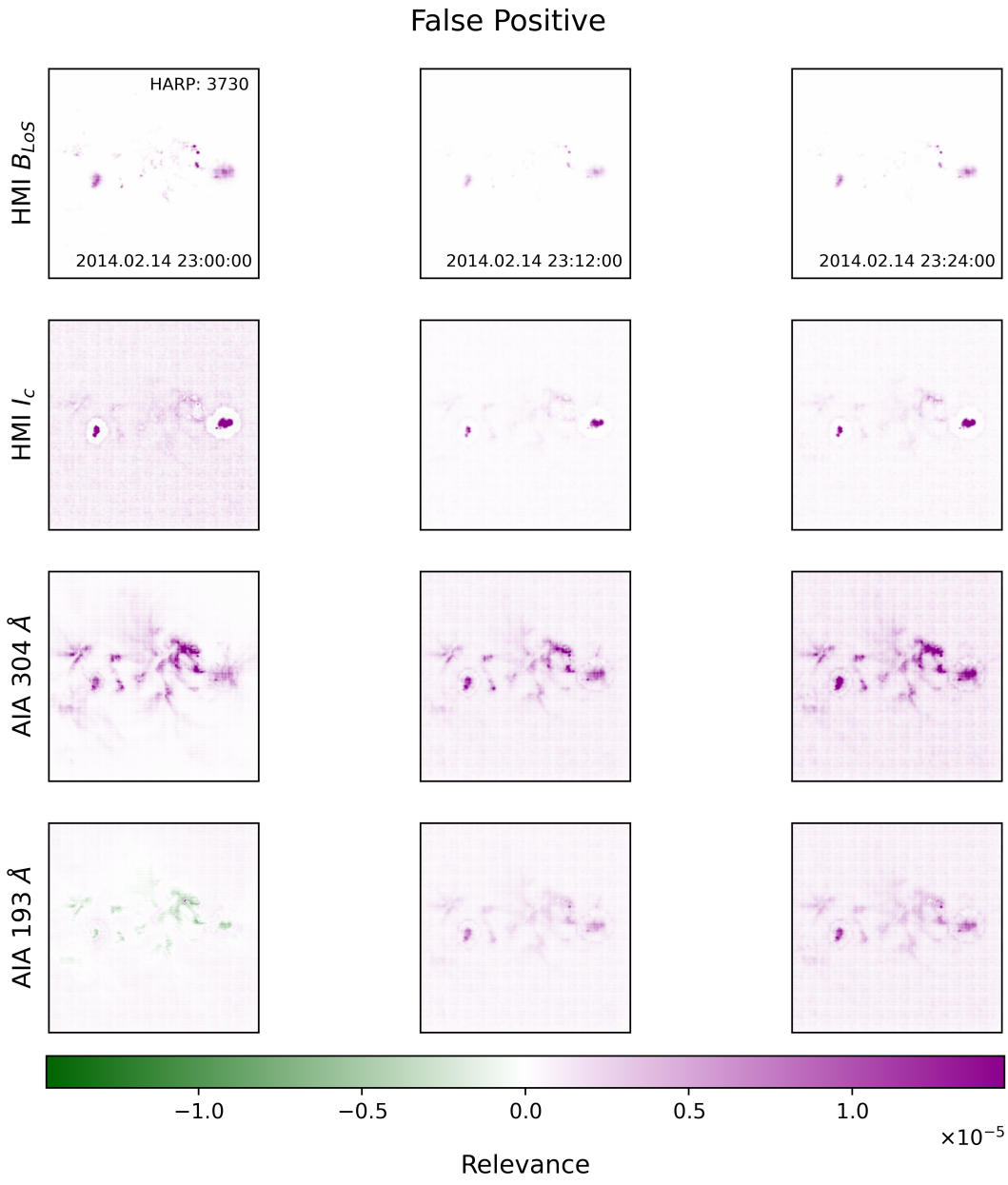


FIGURE 7.7: LRP relevance maps for the *FP* case, HARP 3730. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å.

shown in Figure 7.4. Compared with the HMI relevance, the AIA 193 Å and AIA 304 Å relevance maps show higher values and are spatially associated with coronal and transition–region emission; however, the relevance appears rather spread over the background.

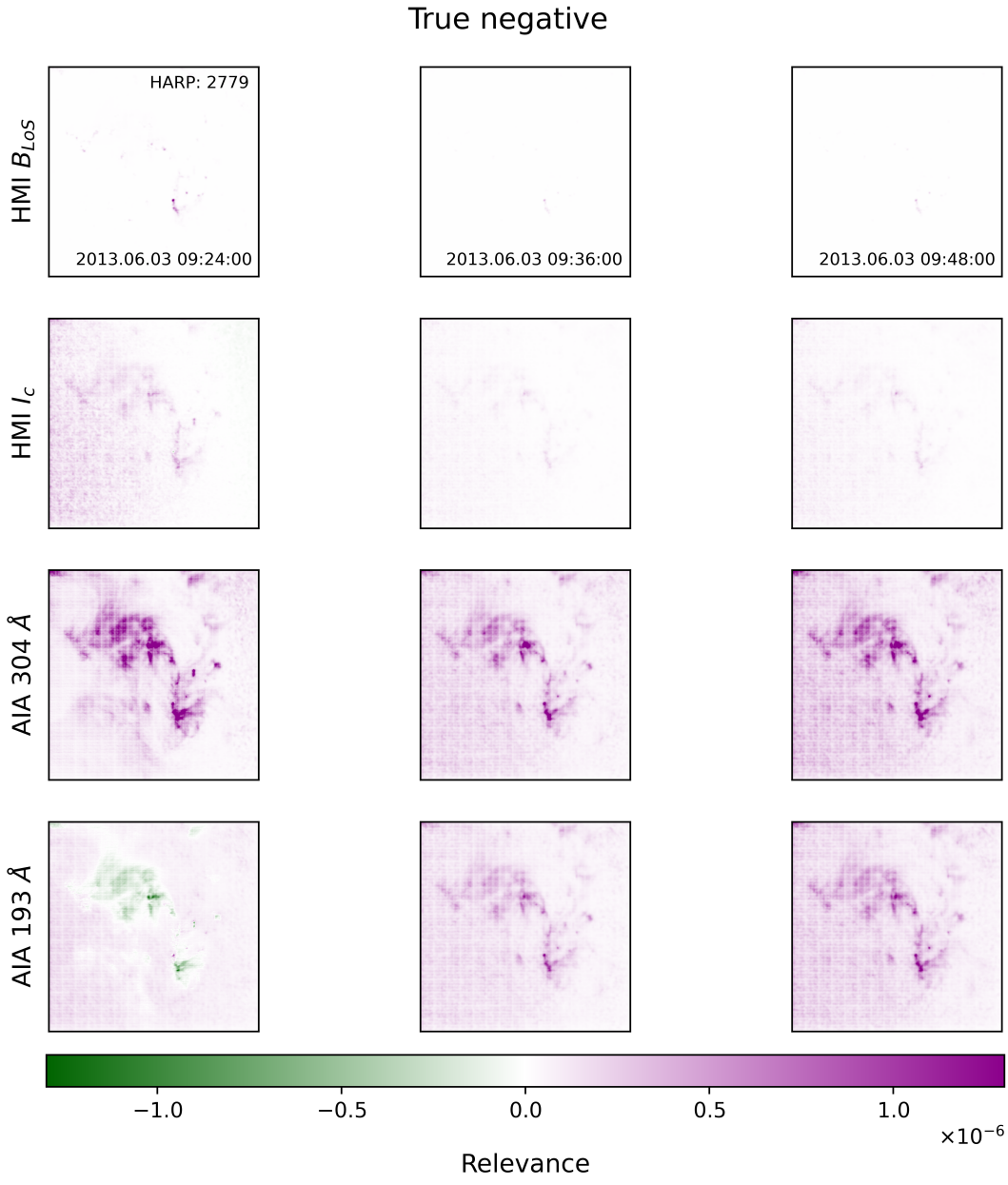


FIGURE 7.8: LRP relevance maps for the *TN* case, HARP 2779. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å.

The corresponding last–time–step input images in Figure 7.4 and physics–based diagnostic maps in Figure 7.5 indicate that this is a relatively simple AR. The HMI images show very limited magnetic structure and no compact sunspot system, while the diagnostic maps show no clear nonpotential signatures, such as localized enhancements of free magnetic energy density, coherent current–helicity structures, or strong magnetic–field gradients. This

supports the interpretation that a diffuse relevance pattern is associated with an AR lacking localized photospheric and magnetic features that the model could use as salient flare–related evidence.

False Negative: HARP 2809

The *FN* case corresponds to HARP 2809, from 2013.05.31 17:36 to 18:00, and is associated with an M1.0 flare, but was classified as non–flaring by all three models, with predicted flare probabilities of $p_M = 0.101$, $p_{MCA} = 0.143$, and $p_{MCAF} = 0.029$. The relevance maps in Figure 7.9 show positive relevance mainly in the AIA channels. The AIA 304 Å relevance is the most visually prominent, with high positive relevance concentrated around bright transition–region emission. The AIA 193 Å relevance shows a similar but less compact pattern around coronal emission, and some loop–like structures are faintly highlighted as relevant. By contrast, the HMI B_{LoS} and I_c relevance are much weaker and only faintly localized around the central photospheric structures, especially near the weak polarity–inversion–line–like structure visible in Figure 7.4.

The final input frame in Figure 7.4 and the physics–based diagnostic maps in Figure 7.5 suggest that this is a more difficult flaring AR to classify than the *TP* case. The AIA channels show compact bright emission, but the HMI channels do not show a large, compact sunspot system or a strongly organized photospheric structure comparable to HARP 5298. The diagnostic maps show low free magnetic energy density, but localized current–helicity structure and magnetic–field gradients around the AR core were present. This suggests that flare–related signatures were present, but confined to a relatively small portion of the AR and not accompanied by the same large–scale, visually prominent photospheric structure and magnetic complexity observed in the correctly detected flare. As in the *TN* case, the photospheric channels receive very little relevance compared with the atmospheric channels. This case therefore illustrates a missed–event mechanism in which flare–relevant signatures are present but compact, localized, and insufficiently emphasized by the learned representation.

7.4.4 Summary of Key Findings

The results presented in this section provide a complementary view of the behavior of the MCAF model. Rather than assessing predictive performance, the analysis examined how relevance is fractionally allocated across input modalities, physical–context variables, time steps, image channels, and classification outcomes. The case–study analysis then complemented these aggregate results by examining the spatial organization of the relevance patterns in representative *TP*, *FP*, *TN*, and *FN* examples. Four main findings emerge.

First, correctly detected flaring samples are predominantly image–driven. For *TP* cases, a larger fraction of total absolute relevance is assigned to the image branch than to the physical–context branch. This suggests that, when the

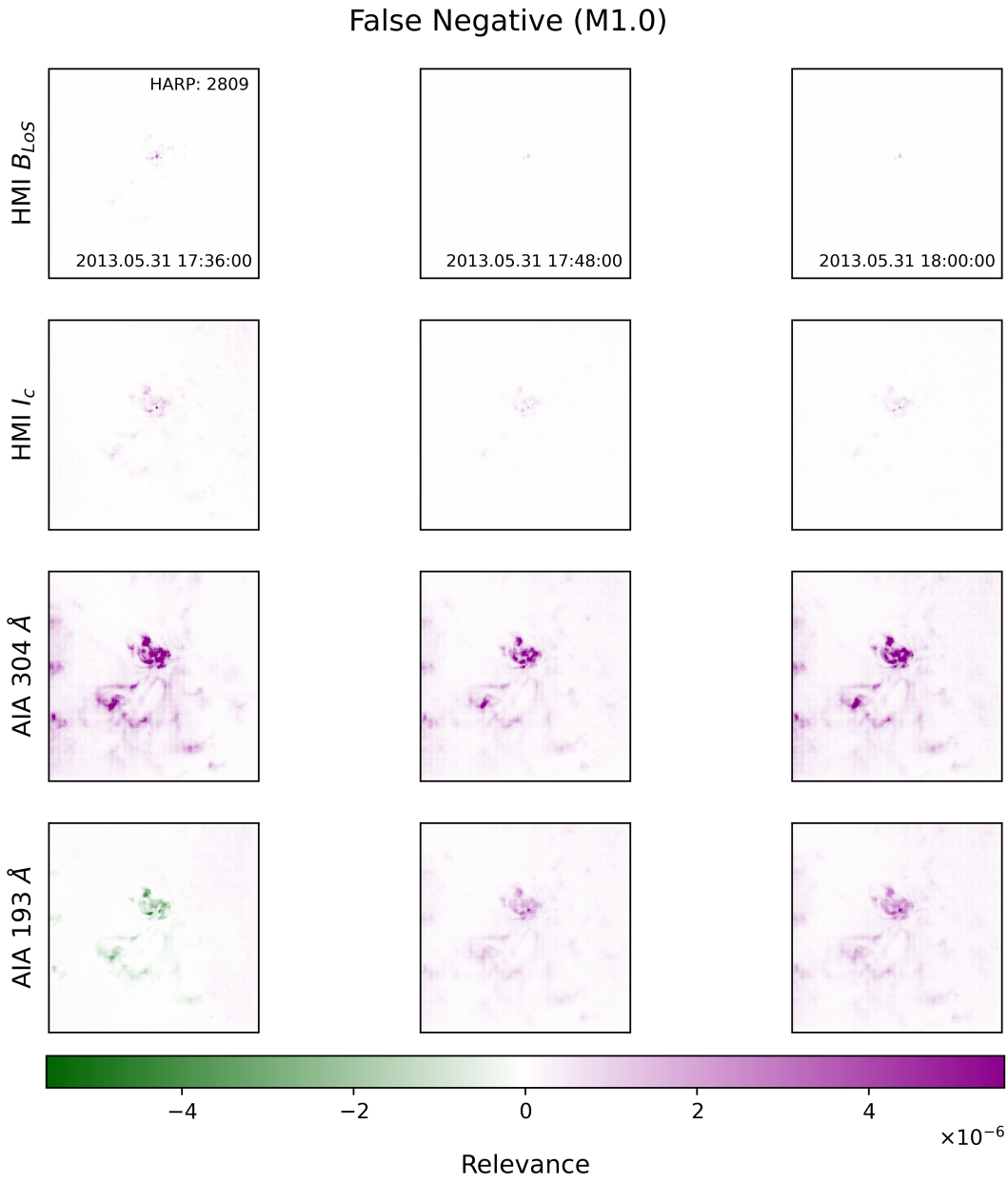


FIGURE 7.9: LRP relevance maps for the *FN* case, HARP 2809, associated with an M1.0 flare. The columns correspond to the three input time steps, and the rows show HMI B_{LoS}, HMI I_c, AIA 193 Å, and AIA 304 Å.

model correctly identifies flare–producing ARs, the decision is mainly supported by visual spatio–temporal information. The physical–parameter distributions show that *TP* cases are generally associated with above–average total unsigned magnetic flux, Φ_{tot}, and total unsigned current helicity, $h_{c,\text{tot}}$. Thus, correctly detected flares are associated with strong image–based evidence of a complex, highly stressed magnetic field.

Second, correctly classified quiet samples show a more balanced relevance allocation. For *TN* cases, the image and physical–context relevance fractions are closer to each other, and their interquartile ranges overlap substantially. This suggests that quiet–region classifications are not explained by a single, uniform attribution pattern. Instead, the model may reject flare productivity using different combinations of image–based information and scalar magnetic descriptors, depending on the structure of the AR. Within the physical–context branch, *TN* cases show a slight shift toward current helicity, suggesting that, when the model relies more strongly on physical parameters, its relevance allocation is biased toward $h_{c,\text{tot}}$.

Third, the channel–wise and temporal relevance analyses show differences in the way the model uses different inputs. At the aggregate test–set level, image relevance is weighted more strongly toward the EUV channels, especially AIA 304 Å, than toward the photospheric HMI channels. The temporal allocation also differs between the two input branches. The image branch assigns larger relevance to the first and last frames of the sequence, whereas the physical–context branch assigns the largest relevance to the central time step. This suggests that the image branch may emphasize changes between the boundary frames, while the physical–parameter time series provides a more representative short–term magnetic state.

Fourth, the misclassified cases suggest two different error regimes, although these results should be interpreted only descriptively because of the small number of *FP* and *FN* samples. *FP* cases show an image–dominant relevance allocation, suggesting that false alarms may occur when active–looking image morphology is interpreted as flare–relevant. *FN* cases show a stronger relative contribution from the physical–context branch, suggesting that missed flares may occur when flare–related image signatures are compact, localized, or insufficiently emphasized by the learned representation. The standardized physical–parameter distributions are consistent with this interpretation: the misclassified cases occupy intermediate regimes between the clearly flaring and clearly quiet populations.

The case studies qualitatively support these aggregate findings. The *TP* example shows coherent relevance patterns associated with a magnetically complex AR, while the *TN* example shows diffuse relevance in a relatively simple AR with a less organized magnetic field. The *FP* example indicates that active–looking morphology can support a positive prediction even when the physics–based diagnostic maps show a less organized magnetic configuration than in the *TP* case. The *FN* example illustrates a missed–event scenario in which flare–related signatures are present but compact and localized. These comparisons with the physics–based diagnostic maps are qualitative rather than pixel–by–pixel, because the physics–based diagnostic maps are

computed in CEA coordinates whereas the model inputs are kept in CCD coordinates. Overall, the results suggest that the MCAF model bases its predictions on a combination of atmospheric image structure and global magnetic context, and that its relevance patterns are broadly consistent with physically meaningful AR structures.

7.5 Conclusions

In this chapter, we extended the multimodal flare–forecasting analysis of Chapter 6 by moving from prediction to interpretation. Using LRP, we examined how the modified MCAF model allocates relevance across the image branch, within the physical–context branch, temporal steps, image channels, and classification outcomes. The goal was to assess whether the model decisions are supported by physically meaningful information and to identify attribution patterns associated with successful and failed predictions.

The analysis shows that the model does not rely uniformly on all input sources. Correctly detected flaring samples are mainly supported by image–based information, especially by EUV channels, while the physical–context branch provides complementary information related to global magnetic properties. Correctly classified quiet samples show a more balanced relevance allocation, suggesting that quiet–region classifications may depend on different combinations of image–based information and scalar magnetic descriptors. The standardized physical–parameter distributions further show that correctly detected flaring ARs are generally associated with above–average Φ_{tot} and $h_{c,\text{tot}}$, whereas correctly classified quiet ARs are associated with lower values of both quantities.

Our results are consistent with previous explainable artificial intelligence studies in solar flare forecasting, which have shown that deep–learning models can learn active–region–related features from solar observations. For example, Sun et al. (2022) showed that Integrated Gradients (IG) could attribute increases in convolutional neural network (CNN) flare predictions to the emergence of magnetic flux that outweighed magnetic–flux cancellation. The authors also showed that Gradient–weighted Class Activation Mapping (Grad–CAM) could roughly reveal physically meaningful flare–related features, particularly strong magnetic fields and PILs. The work of Sun et al. (2022) represented an improvement over earlier attribution–based studies, such as Bhattacharjee et al. (2020) and Yi et al. (2021), by providing higher–resolution, pixel–level explanations that enabled a more detailed interpretation of flare–related magnetic features. Similarly, Pandey, Angryk, and Aydin (2023) and Pandey et al. (2023) applied different post hoc attribution and attention–based interpretation methods to full–disk CNN flare–forecasting models; although these models predict whether the Sun will produce a $\geq$M–class flare within a given forecasting window rather than the exact flare location on the solar disk, the attribution analyses showed that their predictions are driven by features associated with ARs. More recently, Adeyeha, Pandey, and Aydin (2024) introduced a proximity measure for evaluating

whether attribution maps from full–disk CNN flare–forecasting models spatially align with observed ARs and flare locations, providing quantitative interpretability metrics beyond visual inspection. The present work extends this interpretability analysis to a time–evolving multimodal setting, in which the model is interpreted not only through spatial relevance maps, but also through test–set aggregated quantitative measures of how the relevance budget is allocated between image–based information and physical–context variables.

The case studies provide qualitative support for these aggregate results. The *TP* example shows coherent relevance patterns in a magnetically complex AR, whereas the *TN* example shows diffuse relevance in a relatively simple AR with a less organized magnetic field. The *FP* and *FN* examples illustrate two possible failure regimes: false alarms may occur when active–looking morphology is interpreted as flare–relevant, while missed flares may occur when flare–related signatures are compact, localized, or insufficiently emphasized by the learned representation. These results suggest that model errors arise in ambiguous regimes where image morphology and magnetic context do not provide equally clear evidence.

The analysis also clarifies the role and limitations of post hoc interpretability in our setting. The relevance maps should not be interpreted as direct measurements of flare–triggering mechanisms, but as model–specific explanations of how the trained network uses the available inputs. This distinction is important because attribution methods can reveal physically meaningful active–region–related evidence, but they can also expose model shortcuts, sensitivity to observational artifacts, or limitations of the learned representation. In the present analysis, the relevance patterns are broadly consistent with AR morphology and magnetic complexity, but the comparison with the physics–based diagnostic maps remains qualitative. This is because the diagnostic maps are not model inputs and are computed in CEA coordinates, whereas the model inputs are in the original CCD coordinates. Therefore, the main contribution of the interpretability analysis is not to identify causal flare precursors, but to assess whether the model evidence is physically plausible and to diagnose the conditions under which the model may fail.

Finally, the results suggest a limitation of the present multimodal design. The physical–context branch provides global active–region–level information through Φ_{tot} and $h_{c,\text{tot}}$, but it does not provide spatially resolved information about where magnetic stress, current helicity, or strong magnetic gradients are located within the AR. As a result, the model can combine image morphology with global magnetic context, but it cannot explicitly align local image evidence with local magnetic diagnostics. This limitation may contribute to ambiguous cases, such as *FP* samples with active–looking morphology but less organized magnetic diagnostics, or *FN* samples where flare–related signatures are present but compact and localized. Future models could therefore benefit from incorporating spatially resolved and spatially aligned magnetic diagnostics, allowing local image evidence to be compared more directly with local magnetic complexity.

Overall, this chapter shows that interpretability analysis is an essential

complement to predictive evaluation. It helps clarify which input sources contribute to the model decisions, whether the relevance patterns are broadly consistent with physically meaningful AR structures, and under which conditions the model can fail. Within the broader thesis, these results reinforce the central idea that solar flare forecasting benefits from representations that combine spatial structure, temporal evolution, multiwavelength atmospheric information, and physically motivated magnetic context.

Chapter 8

Conclusions

"Ognuno sta solo sul cuor della terra
trafitto da un raggio di sole:
ed è subito sera. "

Salvatore Quasimodo
Ed è subito sera

This chapter concludes the thesis by summarizing the main findings and contributions of the work and outlining future perspectives. The discussion follows the progression of the thesis from representation learning, to temporal forecasting, to multimodal short–term prediction, and finally to model interpretation. In doing so, we emphasize the central argument of the thesis: deep–learning methods can contribute to solar flare forecasting not only by improving predictive performance, but also by learning physically meaningful representations and by interpreting model decisions in relation to the physical properties of flare–productive active regions (ARs).

8.1 Main Findings and Contributions

The first experimental phase of the thesis focused on representation learning of solar ARs. In Chapter 4, a convolutional autoencoder was trained to reconstruct line–of–sight (LoS) magnetograms of ARs. The reconstruction task was not treated as an end in itself, but as an indirect way to evaluate whether the latent representation retained information about the spatial organization of the photospheric magnetic field. Since magnetograms are measurements of a physical field, reconstruction quality was evaluated not only through residual metrics, but also through structural similarity, spectral behavior, pixel–value distributions, and, most importantly, through the temporal evolution of physically derived magnetic quantities.

The results show that the choice of training objective plays a central role not only in determining the quality of the reconstruction, but also in determining the physical consistency of the learned representation. In particular, the finite–difference gradient–regularized objective provided the best overall

performance among the selected models. It improved pixel–wise reconstruction accuracy, structural similarity, and the preservation of spectral power at high values of k. This result is physically meaningful because quantities such as the gradient–weighted neutral–line length and the mean magnetic–field gradient are sensitive to small–scale magnetic structures characterized by sharp magnetic gradients. These features are known indicators of flare–productive ARs and are among the most difficult to reconstruct if the model learns only a smooth visual approximation of the input magnetogram, as occurs when a reconstruction–only objective is used. Consequently, the use of the finite–difference gradient–regularized objective also translated into greater temporal consistency of the physical parameters derived from the reconstructed magnetograms.

The main contribution of this phase was to show that compact neural representations of AR LoS magnetograms can preserve physically meaningful magnetic information, provided that the training objective is designed to constrain not only pixel–wise reconstruction errors, but also discrepancies in local magnetic–field gradients.

Chapter 5 extended the representation–learning framework to sequence–based major–flare forecasting. In this phase, the representations of AR LoS magnetograms produced by the reconstruction–only encoder were organized into temporal sequences and used as input to a recurrent classifier to predict whether an AR would produce at least one M– or X–class flare during a future forecasting window. This step was motivated by the fact that flare productivity depends not only on the instantaneous magnetic configuration of an AR, but also on its evolution. The build–up, persistence, and reorganization of magnetic structures can contain information about the preflare state that is not fully captured by a single static observation.

The results support the use of temporal modeling. The sequence–based framework achieved strong performance for major–flare prediction, demonstrating that the temporal organization of the learned magnetogram representations contains useful predictive information. Among the temporal configurations considered, the combination of a 12 h observational window, a 12 h forecasting window, and static undersampling provided the strongest overall test–set performance, with the best model reaching a *TSS* of 0.892 while maintaining a balanced trade–off between *Recall* and *Specificity*. This suggests that, among the tested configurations, an intermediate temporal window provided enough context to capture relevant preflare magnetic evolution without introducing excessive variability from longer histories or relying only on short–term fluctuations.

The main contribution of this phase was to demonstrate that learned spatial representations of AR LoS magnetograms can be extended to the temporal domain and used effectively for major–flare forecasting.

The main limitation of this phase is that the forecasting model relied exclusively on temporal sequences of encoded LoS magnetograms. Although these representations contain useful information about the evolution of the photospheric magnetic field, they provide only one view of the preflare evolution of an AR. Since flare initiation involves magnetic energy stored in non–potential coronal fields and a coupled response across multiple layers of the

solar atmosphere, a more complete description of the preflare state requires observations beyond LoS magnetograms alone. This limitation motivated the multimodal extension developed in Chapter 6.

Chapter 6 moved from single–modality flare forecasting to short–term multimodal spatio–temporal flare forecasting. The model combined temporally ordered Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI) and SDO/Atmospheric Imaging Assembly (AIA) observations, including LoS magnetograms, continuum intensity images, and extreme–ultraviolet (EUV) images in the 193 Å and 304 Å channels. These image sequences were complemented by scalar physical parameters derived from vector magnetograms, specifically the total unsigned magnetic flux and the total unsigned current helicity. In this way, the model integrated local spatial information learned directly from image data with global physical descriptors related to the size, strength, and non–potentiality of the active–region magnetic field.

The results show a clear improvement as additional sources of information are incorporated. The baseline configuration using only LoS magnetograms already achieved good performance, confirming the result of Chapter 5 that temporal photospheric magnetic information has predictive value. Adding continuum and EUV information further improved the results, indicating that multiwavelength observations provide complementary information about the preflare state of ARs. The best performance was obtained when the physical parameters were also included. In this full multimodal configuration, evaluated on the balanced test set, the model reached an *Accuracy* of 92.74%, a *Recall* of 97.56%, and a *TSS* of 0.86. The improvement in recall is particularly important in the context of flare forecasting, because it indicates that fewer flaring sequences were missed by the model.

The diagnostic analyses support the same conclusion. The predicted–probability distributions showed increasingly clear separation between flaring and nonflaring sequences as the input configuration became richer. The ROC curve analysis confirmed that the multimodal configurations improved threshold–independent discrimination, while the threshold–dependent performance curves showed that the full model maintained strong skill over a broad range of decision thresholds. The calibration analysis indicates that the predicted probabilities were generally informative, although deviations from perfect calibration remained, especially at intermediate probability values. This result is important because operational forecasting requires not only accurate classifications, but also reliable probability estimates that can be interpreted according to different risk tolerances.

The main contribution of this phase was to show that multiwavelength observations and parameters summarizing the magnetic state of the system can be integrated within a common short–term forecasting framework, with the richer input configurations yielding higher performance on the present test set and providing a more complete data–driven description of active–region preflare evolution.

The main limitation of this phase is that, although the multimodal model

improved forecasting performance, suggesting that the additional information was in some way relevant for prediction, the model remained an intrinsically black–box predictor whose decisions are difficult to interpret directly. This limitation motivated the interpretability analysis developed in Chapter 7, where an attribution method was used to examine the model decisions in relation to the model inputs.

Chapter 7 addressed the interpretability of the multimodal forecasting model. This step was essential because the scientific value of a forecasting model depends not only on its predictive skill, but also on whether its predictions can be examined in relation to the physical properties of ARs. The interpretation method selected for this phase was Layer–wise Relevance Propagation (LRP), which consists of propagating the model's output backward through the network to assign relevance to the input variables. This made it possible to analyze the contribution of different image modalities, spatial regions, time steps, and physical parameters to the flare/no–flare decision and to examine whether the model relied on physically meaningful information.

This interpretability step is particularly important because deep–learning models can achieve high performance for reasons that are not necessarily physically meaningful. A model may learn biases related to data selection or correlations that do not directly reflect flare–producing mechanisms. Attribution analysis cannot by itself eliminate this possibility, but it provides a diagnostic tool for examining whether the learned decision process is consistent with the physical interpretation of the problem. In this thesis, interpretability therefore served not as an optional visualization step, but as part of the scientific evaluation of the forecasting model.

The main contribution of this phase was to demonstrate that attribution analysis can be used to examine the internal decision process of the multimodal forecasting model and to assess whether its predictions are supported by features that are physically consistent with known properties of flare–productive ARs.

Generally speaking, attribution methods identify relevance, not causality. Consequently, a region or a global physical descriptor receiving high relevance can be important for the model prediction, but this does not prove that it is the physical cause of the flare. Attribution maps also depend on the architecture, the propagation rules, and the distribution of the training data. Therefore, interpretability results must be understood as model–based evidence rather than direct physical proof. Their value lies in helping to test whether the model behaves consistently with known solar–physics expectations and in identifying cases where the model may rely on unexpected or potentially spurious information.

Taken together, the four experimental phases show that deep learning can contribute to flare forecasting in a physically informative way when the modeling framework is designed around the specific characteristics of the solar flare forecasting problem. The thesis began with the learning of compact representations of photospheric magnetic fields, then used the resulting representations to model active–region evolution, expanded the framework

to include information from multiple atmospheric layers and magnetic summary parameters, and finally interpreted the trained model in terms of relevance patterns. This progression supports the broader implication that flare forecasting should be treated as a spatio–temporal and multimodal problem. ARs are not static entities, and flare occurrence is not determined by a single scalar parameter or a single static observation. Rather, flares arise from the evolving magnetic and atmospheric state of the AR. In this context, deep–learning models are particularly suitable because they can combine different sources of information by design. The results of this thesis show that they are most useful when this flexibility is accompanied by an interpretability analysis that supports scientific understanding.

8.2 Future Perspectives

The results presented in this thesis suggest several directions for future work. A first direction concerns the representation–learning stage. Chapter 4 showed that explicitly constraining local magnetic–field gradients improves the reconstruction of LoS magnetograms and the temporal consistency of the derived physical parameters, particularly for physical parameters that depend on small–scale features and steep gradients. However, the analysis also showed that reconstruction quality is not determined only by pixel–wise residuals or local morphology. Future work could therefore investigate additional terms in the reconstruction loss designed to preserve structure across spatial scales, for example by reducing discrepancies between the Fourier–domain representations of the original and reconstructed LoS magnetograms. Such an approach could allow different spatial frequencies to be weighted differently, or specific spatial scales to be targeted, while preserving the local magnetic morphology needed for physically meaningful active–region representations.

No downstream comparison among the alternative representation–learning objectives evaluated in Chapter 4 was performed. Therefore, future work should assess whether the improvements in reconstruction quality and in the temporal consistency of the derived magnetic–field parameters observed with the gradient–regularized objectives translate into improved forecasting performance.

A related direction concerns how representation learning is integrated with the downstream forecasting task. In this thesis, the image encoders were trained separately and then kept fixed and used as feature extractors for the forecasting models, allowing the representation–learning and forecasting stages to be separated while reducing the computational cost of the supervised training stage. However, no direct comparison with end–to–end training or fine–tuning of the encoders was performed. Therefore, apart from the computational advantage, no additional benefit of this training strategy can be established from the present results. Future work should compare these training strategies to determine whether adapting the encoders directly to the forecasting objective provides additional predictive benefit.

A second direction concerns the temporal formulation of the forecasting problem. Chapter 5 showed that the temporal organization of learned magnetogram representations contains useful information for major–flare forecasting, but the optimal relation between observational–window length and forecasting–window length remains an open question. Future work could therefore perform a more systematic temporal sensitivity analysis. One study could fix the forecasting window and vary the observational–window length, in order to identify how much temporal context is most informative for the model. A complementary study could fix the observational–window length and vary the forecasting window, in order to determine how predictive performance changes with the forecasting horizon. This would help separate the effect of available temporal context from the intrinsic difficulty of forecasting flares over longer lead times.

A third direction is to move toward more operationally realistic forecasting tasks. In this thesis, the goal was to distinguish between flaring and quiet active–region sequences and, to this end, a relatively simplified labeling strategy was introduced to produce a clearer separation between the two classes. In particular, the construction of the negative class excluded sequences with decaying flare activity and sequences leading to flares below the chosen threshold. Future work should therefore test the model under broader task definitions. Including the recent flaring history of each AR as an additional input could provide useful context for distinguishing truly quiet regions from regions that remain flare–capable. This extension would make the forecasting problem closer to the conditions encountered in operational settings.

A related issue concerns class imbalance. Balanced datasets are useful for evaluating the intrinsic discriminative capability of a model, but real flare forecasting is characterized by a strong imbalance between flaring and nonflaring examples. Future work should therefore evaluate the proposed models on datasets with realistic class distributions and compare different imbalance–handling strategies. This should be accompanied by a careful analysis of decision thresholds, since the preferred operating point depends on the relative cost of missed flares and false alarms. The calibration analyses in Chapters 5 and 6 showed that classification skill and probabilistic reliability do not necessarily coincide. Post hoc calibration methods could therefore be useful when the predicted probabilities are intended to support operational decision–making.

Future work should also consider practical extensions of the multimodal framework. In particular, the model could be evaluated on a larger dataset, ideally including observations from multiple solar cycles, and the effect of longer temporal sequences could be tested. These extensions could provide the model with a richer description of active–region evolution and more varied training examples, but they would also increase computational cost, memory requirements, preprocessing demands, and constraints on batch size and model architecture. Therefore, future work should assess whether such extensions provide additional predictive information that justifies the higher computational complexity.

More generally, future work could complement the reported point–estimate comparisons with systematic uncertainty quantification across the experimental phases. This would help assess the robustness of the results.

Overall, future developments should aim to make flare–forecasting models more physically consistent, operationally realistic, and probabilistically reliable. The directions outlined above address complementary aspects of this goal: improving the preservation of magnetic structure in learned representations, clarifying the role of temporal context and forecasting horizon, testing the models under broader and more realistic task definitions, accounting for class imbalance, improving probability calibration, and, particularly in the multimodal setting, evaluating whether larger datasets and longer temporal sequences provide additional predictive value. Together, these extensions would move the work presented in this thesis beyond the controlled experimental setting toward forecasting models that are not only accurate, but also better suited to the complexity of operational solar–flare forecasting.

8.3 Final Remarks

This thesis investigated solar flare forecasting as a task that requires not only accurate prediction, but also physical interpretation. The results show that deep–learning models can learn compact representations of active–region magnetic structure, use the temporal evolution of such representations for major–flare forecasting, improve predictive skill by combining multimodal solar observations with global magnetic descriptors, and provide interpretable relevance patterns that can be compared with known properties of flare–productive ARs.

The main conclusion is therefore not only that deep learning can improve flare prediction, but that it can do so in a way that remains connected to solar physics. Solar flare forecasting remains difficult because flare initiation is governed by complex, nonlinearly evolving magnetic conditions, while direct observational access to the coronal magnetic field, which is essential for understanding this process, remains limited. Nevertheless, the work presented in this thesis shows that physically informed deep–learning approaches, accompanied by interpretation techniques, can contribute to this challenge. By linking representation learning, temporal forecasting, multimodal modeling, and attribution analysis, the thesis provides a step toward forecasting models that are not only accurate, but also scientifically meaningful.

Appendix A

Data Availability

All image data products analyzed in this work are publicly available through the Joint Science Operations Center (JSOC) at http://jsoc.stanford.edu/. Data retrieval was performed using the open-source SunPy library (The SunPy Community et al., 2020).

The list of solar flare events (class C5.0 and above) used to construct the dataset was compiled by Li et al. (2021) from the Geostationary Operational Environmental Satellite (GOES) flare catalog.

The curated dataset defining the positive and negative samples used in this study is publicly available via Zenodo at (Doria Rosales et al., 2025).

Appendix B

Temporal Distribution of the Datasets

This appendix provides a complementary description of the temporal distribution of the datasets used in the experimental phases of the thesis. Chapters 4 and 5 use an active–region–centered dataset, whereas Chapter 6 uses an event–centered dataset. Details of the dataset construction are provided in the corresponding sections of each chapter. In both cases, the training, validation, and test partition is performed at the active–region level, which means that all sequences associated with a given AR are assigned exclusively to a single split, preventing the same AR from contributing samples to different subsets.

B.1 Active–Region–Centered Dataset: Chapters 4 and 5

The dataset used in Chapters 4 and 5 contains 800 ARs satisfying the adopted selection criteria. These ARs span from May 2010 to December 2022, covering most of Solar Cycle 24 and the early phase of Solar Cycle 25.

The dataset is divided at the active–region level into 559 HARPs for training, 144 for validation, and 97 for testing. Figure B.1 shows the temporal distribution of the ARs in each partition together with the monthly sunspot number for reference. The three subsets contain ARs distributed throughout the temporal coverage of the dataset.

B.2 Event–Centered Dataset: Chapter 6

The multimodal dataset used in Chapter 6 is constructed at the sequence level from alist of flare events and quiet active–region observations. The final dataset contains 1206 sequences, comprising 562 positive and 644 negative samples distributed across the training, validation, and test partitions.

The retained positive sequences extend from June 2010 to September 2017 and therefore belong entirely to Solar Cycle 24. The negative sequences extend from May 2010 to September 2024 and include observations from both Solar Cycles 24 and 25. Of the 644 negative sequences, 372 belong to Solar Cycle 24 and 272 to Solar Cycle 25.

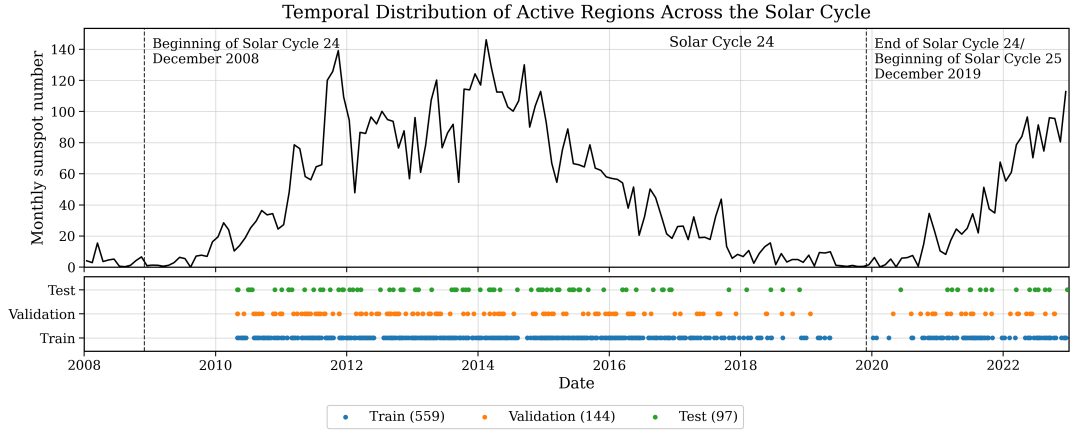


FIGURE B.1: Temporal distribution of the ARs used in Chapters 4 and 5 across the training, validation, and test partitions. The upper panel shows the monthly sunspot number, while the lower panel shows the observation epoch of each AR. The dashed vertical lines indicate the approximate boundaries of Solar Cycles 24 and 25.

Figure B.2 shows the temporal distribution of the positive and negative sequences in each data partition. Marker shape distinguishes the two classes, while color identifies the training, validation, and test subsets.

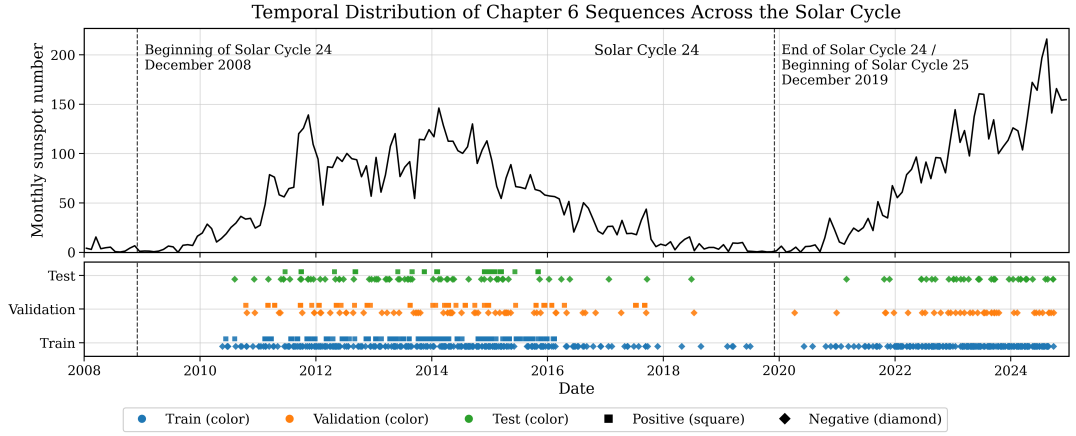


FIGURE B.2: Temporal distribution of the Chapter 6 positive and negative sequences across the training, validation, and test partitions. Marker shape distinguishes positive and negative samples, while color indicates the data partition. The upper panel shows the monthly sunspot number for temporal reference.

Bibliography

Abramenko, Valentyna and Vasyl Yurchyshyn (2010). "Intermittency and multifractality spectra of the magnetic field in solar active regions". In: *The Astrophysical Journal* 722.1, pp. 122–130.

Abramenko, VI (2005). "Relationship between magnetic power spectrum and flare productivity in solar active regions". In: *The Astrophysical Journal* 629.2, pp. 1141–1149.

Abramenko, VI, Tongjiang Wang, and VB Yurchishin (1996). "Analysis of electric current helicity in active regions on the basis of vector magnetograms". In: *Solar Physics* 168.1, pp. 75–89.

Abramenko, VI, VB Yurchishin, and Vincenzo Carbone (1998). "Does the photospheric current take part in the flaring process?" In: *Astronomy and Astrophysics, v. 334, p. L57-L60 (1998)* 334, pp. L57–L60.

Abramenko, VI et al. (2002). "Scaling behavior of structure functions of the longitudinal magnetic field in active regions on the Sun". In: *The Astrophysical Journal* 577.1, pp. 487–495.

— (2003). "Signature of an avalanche in solar flares as measured by photospheric magnetic fields". In: *The Astrophysical Journal* 597.2, pp. 1135–1144.

Adeyeha, Temitope, Chetraj Pandey, and Berkay Aydin (2024). "Large Scale Evaluation of Deep Learning-based Explainable Solar Flare Forecasting Models with Attribution-based Proximity Analysis". In: *2024 IEEE International Conference on Big Data (BigData)*. IEEE, pp. 1209–1214.

Ahmadzadeh, Azim et al. (2021). "How to train your flare prediction model: Revisiting robust sampling of rare events". In: *The Astrophysical Journal Supplement Series* 254.2, p. 23.

Alayrac, Jean-Baptiste et al. (2022). "Flamingo: a visual language model for few-shot learning". In: *Advances in neural information processing systems* 35, pp. 23716–23736.

Alexander, David et al. (1998). "Relative timing of soft X-ray nonthermal line broadening and hard X-ray emission in solar flares". In: *The Astrophysical Journal Letters* 494.2, pp. L235–L238.

Angryk, Rafal A et al. (2020). "Multivariate time series dataset for space weather data analytics". In: *Scientific data* 7.1, p. 227.

Auslender, Alfred and Marc Teboulle (2004). "Interior gradient and epsilon-subgradient descent methods for constrained convex minimization". In: *Mathematics of Operations Research* 29.1, pp. 1–26.

Bach, Sebastian et al. (2015). "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7, e0130140.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473*.

Bamba, Yumi and Kanya Kusano (2018). "Evaluation of applicability of a flare trigger model based on a comparison of geometric structures". In: *The Astrophysical Journal* 856.1, p. 43.

Bamba, Yumi, TT Yamamoto, and TJ Okamoto (2013). "Study on the triggering process of solar flares based on Hinode/SOT observations". In: *The Astrophysical Journal* 778.1, p. 48.

Bamba, Yumi et al. (2017). "Triggering process of the X1. 0 three-ribbon flare in the great active region NOAA 12192". In: *The Astrophysical Journal* 838.2, p. 134.

Barlow, William Henry (1849). "VI. On the spontaneous electrical currents observed in the wires of the electric telegraph". In: *Philosophical Transactions of the Royal Society of London* 139, pp. 61–72.

Barnes, Graham et al. (2016). "A comparison of flare forecasting methods. I. Results from the "all-clear" workshop". In: *The Astrophysical Journal* 829.2, p. 89.

Barta, Veronika et al. (2019). "Effects of solar flares on the ionosphere as shown by the dynamics of ionograms recorded in Europe and South Africa". In: *Annales Geophysicae*. Vol. 37. 4. Copernicus Publications Göttingen, Germany, pp. 747–761.

Barta, Veronika et al. (2022). "Multi-instrumental investigation of the solar flares impact on the ionosphere on 05–06 December 2006". In: *Frontiers in Environmental Science* 10, p. 904335.

Bell, Barbara and Harold Glazer (1959). "Some sunspot and flare statistics". In: *Smithsonian Contribution to Astrophysics, Vol. 3, p. 25* 3, p. 25.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013). "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8, pp. 1798–1828.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE transactions on neural networks* 5.2, pp. 157–166.

Benz, Arnold O (2017). "Flare observations". In: *Living reviews in solar physics* 14.1, p. 2.

Bhattacharjee, Shamik et al. (2020). "Supervised convolutional neural networks for classification of flaring and nonflaring active regions using line-of-sight magnetograms". In: *The Astrophysical Journal* 898.2, p. 98.

Biesecker, DA and BJ Thompson (2000). "Sympathetic flaring with BATSE, GOES, and EIT data". In: *Journal of Atmospheric and Solar-Terrestrial Physics* 62.16, pp. 1449–1455.

Bishop, Christopher M (1995). *Neural networks for pattern recognition*. Oxford university press.

Bishop, Christopher M and Nasser M Nasrabadi (2006). *Pattern recognition and machine learning*. Vol. 4. 4. Springer.

Bobra, Monica G and Sebastien Couvidat (2015). "Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm". In: *The Astrophysical Journal* 798.2, p. 135.

Bobra, Monica G., Xudong Sun, and Michael J. Turmon (2021). *mbobra/SHARPs: SHARPs 0.1.0*. Version v0.1.0. DOI: 10.5281/zenodo.5131292.

Bobra, Monica G et al. (2014). "The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs–space-weather HMI active region patches". In: *Solar Physics* 289.9, pp. 3549–3578.

Boerner, PF et al. (2014). "Photometric and thermal cross-calibration of solar EUV instruments". In: *Solar Physics* 289.6, pp. 2377–2397.

Bornmann, PL and D Shaw (1994). "Flare rates and the McIntosh active-region classifications". In: *Solar physics* 150.1, pp. 127–146.

Borrero, JM et al. (2011). "VFISV: very fast inversion of the Stokes vector for the helioseismic and magnetic imager". In: *Solar Physics* 273.1, pp. 267–293.

Bottou, Léon (2010). "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*. Springer, pp. 177–186.

Box, George EP et al. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brier, Glenn W. et al. (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Buonsanto, M Ji (1999). "Ionospheric storms—a review". In: *Space Science Reviews* 88.3, pp. 563–601.

Buzás, Attila et al. (2023). "Investigating the effect of large solar flares on the ionosphere based on novel Digisonde data comparing three different methods". In: *Frontiers in Astronomy and Space Sciences* 10, p. 1201625.

Cadavid, AC et al. (1994). "Multifractal models of small-scale solar magnetic fields". In: *The Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 429, no. 1, p. 391-399* 429, pp. 391–399.

Campi, Cristina et al. (2019). "Feature ranking of active region source properties in solar flare forecasting and the uncompromised stochasticity of flare occurrence". In: *The Astrophysical Journal* 883.2, p. 150.

Camporeale, Enrico (2019). "The challenge of machine learning in space weather: Nowcasting and forecasting". In: *Space weather* 17.8, pp. 1166–1207.

Camporeale, Enrico and Thomas E Berger (2025). "Verification of the NOAA Space Weather Prediction Center solar flare forecast (1998–2024)". In: *Space Weather* 23.10, e2025SW004546.

Canfield, Richard C, Hugh S Hudson, and David E McKenzie (1999). "Sigmoidal morphology and eruptive solar activity". In: *Geophysical Research Letters* 26.6, pp. 627–630.

Canfield, Richard C, Hugh S Hudson, and Alexei A Pevtsov (2000). "Sigmoids as precursors of solar eruptions". In: *IEEE transactions on plasma science* 28.6, pp. 1786–1794.

Cannon, Paul S et al. (2013). "Extreme space weather: Impacts on engineered systems-a summary". In.

Carion, Nicolas et al. (2020). "End-to-end object detection with transformers". In: *European conference on computer vision*. Springer, pp. 213–229.

Carrington, Richard Christopher (1858). "On the distribution of the solar spots in latitudes since the beginning of the year 1854, with a map". In: *Monthly Notices of the Royal Astronomical Society, Vol. 19, p. 1-3* 19, pp. 1–3.

Cauchy, Augustin et al. (1847). "Méthode générale pour la résolution des systemes d'équations simultanées". In: *Comp. Rend. Sci. Paris* 25.1847, pp. 536–538.

Centeno, Rebecca et al. (2014). "The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: optimization of the spectral line inversion code". In: *Solar Physics* 289.9, pp. 3531–3547.

Chaudhari, Anurag et al. (2024). "Association of solar flares with magnetic complexity of the sunspot groups in solar active regions during solar cycles 23–25: A Chaudhari et al." In: *Indian Journal of Physics* 98.9, pp. 3075–3082.

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chen, Guang-ming et al. (2012). "A comparison of the effects of CIR-and CME-induced geomagnetic activity on thermospheric densities and spacecraft orbits: Case studies". In: *Journal of Geophysical Research: Space Physics* 117.A8.

Chen, Yang et al. (2019). "Identifying solar flare precursors using time series of SDO/HMI images and SHARP parameters". In: *Space weather* 17.10, pp. 1404–1426.

Cheng, X et al. (2012). "Differential emission measure analysis of multiple structural components of coronal mass ejections in the inner corona". In: *The Astrophysical Journal* 761.1, p. 62.

Cheung, Mark CM et al. (2015). "Thermal diagnostics with the atmospheric imaging assembly on board the solar dynamics observatory: a validated method for differential emission measure inversions". In: *The Astrophysical Journal* 807.2, p. 143.

Chintzoglou, Georgios and Jie Zhang (2013). "Reconstructing the subsurface three-dimensional magnetic structure of a solar active region using SDO/HMI observations". In: *The Astrophysical Journal Letters* 764.1, p. L3.

Cortie, AL (1901). "On the types of sun-spot disturbances". In: *Astrophysical Journal, vol. 13, p. 260* 13, p. 260.

Couvidat, Sebastien et al. (2016). "Observables processing for the helioseismic and magnetic imager instrument on the solar dynamics observatory". In: *Solar Physics* 291.7, pp. 1887–1938.

Cui, Yanmei et al. (2006). "Correlation between solar flare productivity and photospheric magnetic field properties: 1. Maximum horizontal gradient, length of neutral line, number of singular points". In: *Solar Physics* 237.1, pp. 45–59.

Dalmasse, Kevin et al. (2015). "The origin of net electric currents in solar active regions". In: *The Astrophysical Journal* 810.1, p. 17.

Davies, Kenneth (1990). *Ionospheric radio*. 31. Iet.

Davis, Jesse and Mark Goadrich (2006). "The relationship between Precision-Recall and ROC curves". In: *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240.

Deng, Zheng et al. (2021). "Fine-grained solar flare forecasting based on the hybrid convolutional neural networks". In: *The Astrophysical Journal* 922.2, p. 232.

Devlin, Jacob et al. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.

Dissauer, Karin, KD Leka, and Eric L Wagner (2023). "Properties of Flare-imminent versus Flare-quiet Active Regions from the Chromosphere through the Corona. I. Introduction of the AIA Active Region Patches (AARPs)". In: *The Astrophysical Journal* 942.2, p. 83.

Doornbos, E and H Klinkrad (2006). "Modelling of space weather effects on satellite drag". In: *Advances in Space Research* 37.6, pp. 1229–1239.

Doria Rosales, Elizabeth et al. (Dec. 2025). *Solar flare forecasting dataset*. Version v1.0.0. DOI: 10.5281/zenodo.18018287. URL: https://doi.org/10.5281/zenodo.18018287.

Dosovitskiy, Alexey et al. (2020). "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929*.

Duan, Aiying, Yaoyu Xing, and Chaowei Jiang (2024). "Are Solar Active Regions Born with Neutralized Currents?" In: *Research in Astronomy and Astrophysics* 24.7, p. 075005.

Dumoulin, Vincent and Francesco Visin (2016). "A guide to convolution arithmetic for deep learning". In: *arXiv preprint arXiv:1603.07285*.

Dyer, C et al. (2000). "Cosmic radiation effects on avionics, an increasing hazard in the new millennium". In: *22nd International Congress on Aeronautical Sciences, Harrogate, UK*.

Elman, Jeffrey L (1990). "Finding structure in time". In: *Cognitive science* 14.2, pp. 179–211.

Esser, Patrick, Robin Rombach, and Bjorn Ommer (2021). "Taming transformers for high-resolution image synthesis". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Fletcher, Lyndsay et al. (2011). "An observational overview of solar flares". In: *Space science reviews* 159.1, p. 19.

Forbes, TG (2000). "A review on the genesis of coronal mass ejections". In: *Journal of Geophysical Research: Space Physics* 105.A10, pp. 23153–23165.

Gallagher, Peter T, Y-J Moon, and Haimin Wang (2002). "Active-region monitoring and flare forecasting–i. data processing and first results". In: *Solar Physics* 209.1, pp. 171–183.

George, PL and PA Bradley (1973). "Relationship between HF absorption at vertical and oblique incidence". In: *Proceedings of the Institution of Electrical Engineers*. Vol. 120. 11. IET, pp. 1355–1361.

Georgoulis, Manolis K (2005). "Turbulence in the solar atmosphere: Manifestations and diagnostics via solar image processing". In: *Solar Physics* 228.1, pp. 5–27.

— (2012). "Are solar active regions with major flares more fractal, multifractal, or turbulent than others?" In: *Solar Physics* 276.1, pp. 161–181.

Georgoulis, Manolis K, Viacheslav S Titov, and Zoran Mikić (2012). "Non-neutralized electric current patterns in solar active regions: origin of the shear-generating Lorentz force". In: *The Astrophysical Journal* 761.1, p. 61.

Georgoulis, Manolis K et al. (2021). "The flare likelihood and region eruption forecasting (FLARECAST) project: flare forecasting in the big data & machine learning era". In: *Journal of Space Weather and Space Climate* 11, p. 39.

Georgoulis, Manolis K et al. (2024). "Prediction of solar energetic events impacting space weather conditions". In: *Advances in Space Research*.

Giovanelli, Ronald Gordon (1939). "The Relations Between Eruptions and Sunspots." In: *Astrophysical Journal, vol. 89, p. 555* 89, p. 555.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep sparse rectifier neural networks". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, pp. 315–323.

Goodfellow, Ian et al. (2016). *Deep learning*. Vol. 1. 2. MIT press Cambridge.

Goodfellow, Ian J et al. (2014). "Generative adversarial nets". In: *Advances in neural information processing systems* 27.

Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton (2013). "Speech recognition with deep recurrent neural networks". In: *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, pp. 6645–6649.

Greatrix, GR (1963). "On the statistical relations between flare intensity and sunspots". In: *Monthly Notices of the Royal Astronomical Society* 126.2, pp. 123–133.

Green, LM and B Kliem (2009). "Flux rope formation preceding coronal mass ejection onset". In: *The Astrophysical Journal* 700.2, pp. L83–L87.

Green, LM, B Kliem, and AJ Wallace (2011). "Photospheric flux cancellation and associated flux rope formation and eruption". In: *Astronomy & Astrophysics* 526, A2.

Greisen, Eric W and Mark R Calabretta (2002). "Representations of world coordinates in FITS". In: *Astronomy & Astrophysics* 395.3, pp. 1061–1075.

Guastavino, Sabrina et al. (2022). "Implementation paradigm for supervised flare forecasting studies: A deep learning application with video data". In: *Astronomy & Astrophysics* 662, A105.

Guerra, Jordan Armando, A Pulkkinen, and Vadim M Uritsky (2015). "Ensemble forecasting of major solar flares: First results". In: *Space weather* 13.10, pp. 626–642.

Guo, Chuan et al. (2017). "On calibration of modern neural networks". In: *International conference on machine learning*. PMLR, pp. 1321–1330.

Hagyard, MJ (1990). "The significance of vector magnetic field measurements". In: *Società Astronomica Italiana, Memorie (ISSN 0037-8720), vol. 61, no. 2, 1990, p. 337-357.* 61, pp. 337–357.

Hagyard, MJ, P Venkatakrishnan, and JB Smith Jr (1990). "Nonpotential magnetic fields at sites of gamma-ray flares". In.

Hagyard, MJ et al. (1984). "A quantitative study relating observed shear in photospheric magnetic fields to repeated flaring". In: *Solar physics* 91.1, pp. 115–126.

Hale, George E (1908). "On the probable existence of a magnetic field in sun-spots". In: *Astrophysical Journal, vol. 28, p. 315* 28, p. 315.

Hale, George E and Seth B Nicholson (1925). "The law of sun-spot polarity". In: *Astrophysical Journal, vol. 62, p. 270* 62, p. 270.

Hale, George E et al. (1919). "The magnetic polarity of sun-spots". In: *Astrophysical Journal, vol. 49, p. 153* 49, p. 153.

Hale, George Ellery and Seth Barnes Nicholson (1938). "Magnetic observations of sunspots, 1917-1924..." In: *Washington*.

Han, Ke et al. (2023). "Research progress on solar flare forecast methods based on data-driven models". In: *Research in Astronomy and Astrophysics* 23.6, p. 065002.

Han, Ke et al. (2025). "Research progress in solar flare prediction methods". In: *Research in Astronomy and Astrophysics*.

Hannah, IG and EP Kontar (2013). "Multi-thermal dynamics and energetics of a coronal mass ejection in the low solar atmosphere". In: *Astronomy & Astrophysics* 553, A10.

Hapgood, Mike (2018). "Chapter 1 - Societal and Economic Importance of Space Weather". In: *Machine Learning Techniques for Space Weather*. Ed. by Enrico Camporeale, Simon Wing, and Jay R. Johnson. Elsevier, pp. 3–26. ISBN: 978-0-12-811788-0. DOI: https://doi.org/10.1016/B978-0-12-811788-0.00001-9. URL: https://www.sciencedirect.com/science/article/pii/B9780128117880000019.

Harra, LK, SA Matthews, and JL Culhane (2001). "Nonthermal velocity evolution in the precursor phase of a solar flare". In: *The Astrophysical Journal Letters* 549.2, pp. L245–L248.

Harra, LK et al. (2009). "Coronal nonthermal velocity following helicity injection before an X-class flare". In: *The Astrophysical Journal* 691.2, pp. L99–L102.

He, Haibo and Edwardo A Garcia (2009). "Learning from imbalanced data". In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.

He, Haibo and Yunqian Ma, eds. (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press. DOI: 10.1002/9781118646106.

He, Kaiming et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hinton, Geoffrey E and Ruslan R Salakhutdinov (2006). "Reducing the dimensionality of data with neural networks". In: *science* 313.5786, pp. 504–507.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780.

Hoeksema, J Todd et al. (2014). "The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: overview and performance". In: *Solar Physics* 289.9, pp. 3483–3530.

Huang, Gao et al. (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Huang, Xin et al. (2018). "Deep learning based solar flare forecasting model. I. Results for line-of-sight magnetograms". In: *The Astrophysical Journal* 856.1, p. 7.

Huber, Peter J (1992). "Robust estimation of a location parameter". In: *Breakthroughs in statistics: Methodology and distribution*. Springer, pp. 492–518.

Hudson, HS et al. (1998). "X-ray coronal changes during halo CMEs". In: *Geophysical research letters* 25.14, pp. 2481–2484.

Illarionov, Egor Aleksandrovich (2025). "Machine learning methods in solar physics". In: *Uspekhi Fizicheskikh Nauk* 195.4, pp. 395–415.

Imada, S et al. (2008). "Non-Gaussian line profiles in a large solar flare observed on 2006 December 13". In: *The Astrophysical Journal Letters* 679.2, pp. L155–L159.

Ioffe, Sergey and Christian Szegedy (2015). "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. pmlr, pp. 448–456.

Jaeggli, Sarah A (2016). "Multi-wavelength study of a delta-spot. I. A region of very strong, horizontal magnetic field". In: *The Astrophysical Journal* 818.1, p. 81.

Jaegle, Andrew et al. (2021). "Perceiver: General perception with iterative attention". In: *International conference on machine learning*. PMLR, pp. 4651–4664.

Janssens, Jan et al. (2025). "Solar flare rates and probabilities based on the McIntosh classification: Impacts of GOES/XRS rescaling and revisited sunspot classifications". In: *Journal of Space Weather and Space Climate* 15, p. 14.

Jing, Ju et al. (2006). "The statistical relationship between the photospheric magnetic parameters and the flare productivity of active regions". In: *The Astrophysical Journal* 644.2, pp. 1273–1277.

Johnson, Justin M and Taghi M Khoshgoftaar (2019). "Survey on deep learning with class imbalance". In: *Journal of big data* 6.1, p. 27.

Jolliffe, Ian T. and David B. Stephenson, eds. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. Chichester, UK: Wiley-Blackwell. DOI: 10.1002/9781119960003.

Jonas, Eric et al. (2018). "Flare prediction using photospheric and coronal image data". In: *Solar Physics* 293.3, p. 48.

Kawate, T and S Imada (2013). "The Relationship between Extreme Ultraviolet Non-thermal Line Broadening and High-energy Particles during Solar Flares". In: *The Astrophysical Journal* 775.2, p. 122.

Khan, Salman H et al. (2017). "Cost-sensitive learning of deep feature representations from imbalanced data". In: *IEEE transactions on neural networks and learning systems* 29.8, pp. 3573–3587.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114*.

Klinkrad, Heiner (1996). "On the use of atmosphere models in re-entry predictions". In: *Environment Modeling for Space-Based Applications*. Vol. 392, p. 287.

Kontogiannis, Ioannis et al. (2017). "Non-neutralized electric currents in solar active regions and flare productivity". In: *Solar Physics* 292.11, p. 159.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25.

Kubo, Masahito et al. (2007). "Hinode observations of a vector magnetic field change associated with a flare on 2006 December 13". In: *Publications of the Astronomical Society of Japan* 59.sp3, S779–S784.

Kullback, Solomon and Richard A Leibler (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Künzel, Horst (1959). "Die Flare-Häufigkeit in Fleckengruppen unterschiedlicher Klasse und magnetischer Struktur (Mitteilungen des Astrophysikalischen Observatoriums Potsdam Nr. 87)". In: *Astronomische Nachrichten* 285.5-6, pp. 271–273.

— (1965). "Zur Klassifikation von Sonnenfleckengruppen". In: *Astronomische Nachrichten, volume 288, p. 177* 288, p. 177.

Kutiev, Ivan et al. (2013). "Solar activity impact on the Earth's upper atmosphere". In: *Journal of Space Weather and Space Climate* 3, A06.

Lawrence, JK, AC Cadavid, and AA Ruzmaikin (1996). "On the multifractal distribution of solar magnetic fields". In: *Astrophysical Journal v. 465, p. 425* 465, p. 425.

Lawrence, JK, AA Ruzmaikin, and AC Cadavid (1993). "Multifractal measure of the solar magnetic field". In: *Astrophysical Journal v. 417, p. 805* 417, p. 805.

LeCun, Yann et al. (2002). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

Lee, Kangjin et al. (2012). "Solar flare occurrence rate and probability in terms of the sunspot classification supplemented with sunspot area and its changes". In: *Solar Physics* 281.2, pp. 639–650.

Lei, Jiuhou et al. (2013). "Impact of solar forcing on thermospheric densities and spacecraft orbits from CHAMP and GRACE". In: *Geodetic Sciences-Observations, Modeling and Applications*. IntechOpen.

Leka, KD and G Barnes (2003a). "Photospheric magnetic field properties of flaring versus flare-quiet active regions. I. Data, general approach, and sample results". In: *The Astrophysical Journal* 595.2, pp. 1277–1295.

Leka, KD and G Barnes (2003b). "Photospheric magnetic field properties of flaring versus flare-quiet active regions. II. Discriminant analysis". In: *The Astrophysical Journal* 595.2, pp. 1296–1306.

— (2007). "Photospheric magnetic field properties of flaring versus flare-quiet active regions. IV. A statistically significant sample". In: *The Astrophysical Journal* 656.2, pp. 1173–1186.

Leka, KD et al. (1996). "Evidence for current-carrying emerging flux". In: *Astrophysical Journal v. 462, p. 547* 462, p. 547.

Leka, KD et al. (2019). "A comparison of flare forecasting methods. II. Benchmarks, metrics, and performance results for operational solar flare forecasting systems". In: *The Astrophysical Journal Supplement Series* 243.2, p. 36.

Lemen, James R et al. (2012). "The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO)". In: *Solar Physics* 275.1, pp. 17–40.

Levine, Sergey et al. (2016). "End-to-end training of deep visuomotor policies". In: *Journal of Machine Learning Research* 17.39, pp. 1–40.

Li, Ting et al. (2021). "Magnetic flux and magnetic nonpotentiality of active regions in eruptive and confined solar flares". In: *The Astrophysical Journal Letters* 917.2, p. L29.

Li, Xuebao et al. (2020). "Predicting solar flares using a novel deep convolutional neural network". In: *The Astrophysical Journal* 891.1, p. 10.

Lim, Eun-Kyung et al. (2010). "The formation of a magnetic channel by the emergence of current-carrying magnetic fields". In: *The Astrophysical Journal* 719.1, pp. 403–414.

Lin, Tsung-Yi et al. (2017). "Focal loss for dense object detection". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988.

Liu, Chang et al. (2017a). "Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm". In: *The Astrophysical Journal* 843.2, p. 104.

Liu, Hao et al. (2019). "Predicting solar flares using a long short-term memory network". In: *The Astrophysical Journal* 877.2, p. 121.

Liu, Jing et al. (2021a). "Electrodynamical coupling of the geospace system during solar flares". In: *Journal of Geophysical Research: Space Physics* 126.1, e2020JA028569.

Liu, Jing et al. (2021b). "Solar flare effects in the Earth's magnetosphere". In: *Nature Physics* 17.7, pp. 807–812.

Liu, Jing et al. (2024). "Energy deposition into the ionosphere during a solar flare with extreme-ultraviolet late phase". In: *The Astrophysical Journal Letters* 963.1, p. L8.

Liu, Yang et al. (1996). "Gradients of the line-of-sight magnetic fields in active region NOAA 6659". In: *Solar Physics* 169.1, pp. 79–89.

Liu, Yang et al. (2017b). "Electric-current neutralization, magnetic shear, and eruptive activity in solar active regions". In: *The Astrophysical Journal Letters* 846.1, p. L6.

Livingston, W et al. (2006). "Sunspots with the strongest magnetic fields". In: *Solar Physics* 239.1, pp. 41–68.

Loshchilov, Ilya and Frank Hutter (2017). "Decoupled weight decay regularization". In: *arXiv preprint arXiv:1711.05101*.

Louis, Rohan Eugene (2025). "A Statistical Study of δ-Sunspots from Solar Cycle 23 to 25". In: *Solar Physics* 300.11, p. 151.

Lü, Yanping, Jingxiu Wang, and Huaning Wang (1993). "Shear angle of magnetic fields". In: *Solar physics* 148.1, pp. 119–132.

Lundberg, Scott M. and Su-In Lee (2017). "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17, 4768—4777. ISBN: 9781510860964.

Luong, Minh-Thang, Hieu Pham, and Christopher D Manning (2015). "Effective approaches to attention-based neural machine translation". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1412–1421.

Marchetti, Francesco et al. (2022). "Score-oriented loss (SOL) functions". In: *Pattern Recognition* 132, p. 108913.

Marqué, Ch, P Lantos, and JP Delaboudiniere (2002). "Multi wavelength investigation of the eruption of a sigmoidal quiescent filament". In: *Astronomy & Astrophysics* 387.1, pp. 317–325.

Martres, M-J et al. (1968). "Étude de la localisation des éruptions dans la structure magnétique évolutive des régions actives solaires". In: *Solar Physics* 5.2, pp. 187–206.

Masci, Jonathan et al. (2011). "Stacked convolutional auto-encoders for hierarchical feature extraction". In: *International conference on artificial neural networks*. Springer, pp. 52–59.

McAteer, RT James, Peter T Gallagher, and Paul A Conlon (2010). "Turbulence, complexity, and solar flares". In: *Advances in Space Research* 45.9, pp. 1067–1074.

McAteer, RT James, Peter T Gallagher, and Jack Ireland (2005). "Statistics of active region complexity: A large-scale fractal dimension survey". In: *The Astrophysical Journal* 631.1, pp. 628–635.

McCloskey, Aoife E, Peter T Gallagher, and D Shaun Bloomfield (2016). "Flaring rates and the evolution of sunspot group McIntosh classifications". In: *Solar Physics* 291.6, pp. 1711–1738.

McIntosh, Patrick S (1990). "The classification of sunspot groups". In: *Solar Physics* 125.2, pp. 251–267.

McNish, AG (1940). "The magnetic storm of March 24, 1940". In: *Terrestrial Magnetism and Atmospheric Electricity* 45.3, pp. 359–364.

Melrose, DB (1991). "Neutralized and unneutralized current patterns in the solar corona". In: *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 381, Nov. 1, 1991, p. 306-312.* 381, pp. 306–312.

— (1995). "Current paths in the corona and energy release in solar flares". In: *Astrophysical Journal v. 451, p. 391* 451, p. 391.

— (1996). "Reply to comments by EN Parker". In: *The Astrophysical Journal* 471.1, pp. 497–500.

Milligan, Ryan O (2011). "Spatially resolved nonthermal line broadening during the impulsive phase of a solar flare". In: *The Astrophysical Journal* 740.2, p. 70.

Milligan, Ryan O and Sarah A McElroy (2013). "Continuum contributions to the SDO/AIA passbands during solar flares". In: *The Astrophysical Journal* 777.1, p. 12.

Minsky, Marvin (1961). "Steps toward artificial intelligence". In: *Proceedings of the IRE* 49.1, pp. 8–30.

Minsky, Marvin and S Papert (1969). *Perceptrons-Expanded Edition: An Introduction to Computational Geometry*.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller (2018). "Methods for interpreting and understanding deep neural networks". In: *Digital signal processing* 73, pp. 1–15.

Montavon, Grégoire et al. (2017). "Explaining nonlinear classification decisions with deep taylor decomposition". In: *Pattern recognition* 65, pp. 211–222.

Moore, Ronald L et al. (2026). "The Making of Delta Sunspots". In: *The Astrophysical Journal* 1000.2, p. 270.

Moreton, GE and AB Severny (1968). "Magnetic fields and flares in the region CMP 20 September 1963". In: *Solar Physics* 3.2, pp. 282–297.

Möstl, Christian et al. (2015). "Strong coronal channelling and interplanetary evolution of a solar storm up to Earth and Mars". In: *Nature Communications* 6.1, p. 7135.

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

Nishimoto, Shohei et al. (2021). "Validation of computed extreme ultraviolet emission spectra during solar flares". In: *Earth, Planets and Space* 73.1, p. 79.

Nishizuka, Naoto et al. (2017). "Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms". In: *The Astrophysical Journal* 835.2, p. 156.

Nishizuka, Naoto et al. (2018). "Deep Flare Net (DeFN) model for solar flare prediction". In: *The Astrophysical Journal* 858.2, p. 113. DOI: 10.3847/1538-4357/aab9a7.

Nishizuka, Naoto et al. (2020). "Reliable probability forecast of solar flares: deep flare net-reliable (DeFN-R)". In: *The Astrophysical Journal* 899.2, p. 150.

Normand, Eugene (1996). "Single-event effects in avionics". In: *IEEE Transactions on nuclear science* 43.2, pp. 461–474.

Nwankwo, Victor UJ, Sandip K Chakrabarti, and Robert S Weigel (2015). "Effects of plasma drag on low Earth orbiting satellites due to solar forcing induced perturbations and heating". In: *Advances in Space Research* 56.1, pp. 47–56.

O'dwyer, B et al. (2010). "SDO/AIA response to coronal hole, quiet Sun, active region, and flare plasma". In: *Astronomy & Astrophysics* 521, A21.

Okamoto, Takenori J and Takashi Sakurai (2018). "Super-strong magnetic field in sunspots". In: *The Astrophysical Journal Letters* 852.1, p. L16.

Pandey, Chetraj, Rafal A Angryk, and Berkay Aydin (2023). "Explaining full-disk deep learning model for solar flare prediction using attribution methods". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 72–89.

Pandey, Chetraj et al. (2023). "Explainable deep learning-based solar flare prediction with post hoc attention for operational forecasting". In: *International Conference on Discovery Science*. Springer, pp. 567–581.

Parker, EN (1996). "Comment on "Current paths in the corona and energy release in solar flares"". In: *The Astrophysical Journal* 471.1, pp. 489–496.

Patterson, A and H Zirin (1981). "Transient magnetic field changes in flares". In: *Astrophysical Journal, Part 2-Letters to the Editor, vol. 243, Jan. 15, 1981, p. L99-L101* 243, pp. L99–L101.

Paula, Eurico R de et al. (2022). "Solar flare and radio burst effects on GNSS signals and the ionosphere during September 2017". In: *Radio Science* 57.10, pp. 1–15.

Pearce, G and RA Harrison (1990). "Sympathetic flaring". In: *Astronomy and Astrophysics (ISSN 0004-6361), vol. 228, no. 2, Feb. 1990, p. 513-516.* 228, pp. 513–516.

Pearson, Karl (1895). "VII. Note on regression and inheritance in the case of two parents". In: *proceedings of the royal society of London* 58.347-352, pp. 240–242.

Pesnell, W Dean, B J Thompson, and PC Chamberlin (2012). "The solar dynamics observatory (SDO)". In: *The solar dynamics observatory*. Springer, pp. 3–15.

Pevtsov, Alexei A (2002). "Active-region filaments and X-ray sigmoids". In: *solar physics* 207.1, pp. 111–123.

Poisson, Mariano et al. (2013). "Study of magnetic flux emergence and related activity in active region NOAA 10314". In: *Advances in Space Research* 51.10, pp. 1834–1841.

Preece, William H (1894). "Earth currents". In: *Nature* 49.1276, pp. 554–554.

Priest, E. and T. Forbes (2002). *Magnetic Reconnection: MHD Theory and Applications*. Cambridge University Press.

PyTorch Contributors (n.d.). *BCEWithLogitsLoss*. https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html. Accessed: 2026-05-07.

Ranns, NDR et al. (2000). "Location of the source of soft X-ray non-thermal line broadenings in a solar flare". In: *Astronomy and Astrophysics, v. 364, p. 859-872 (2000)* 364, pp. 859–872.

Reames, Donald V (1999). "Particle acceleration at the Sun and in the heliosphere". In: *Space Science Reviews* 90.3, pp. 413–491.

— (2013). "The two sources of solar energetic particles". In: *Space Science Reviews* 175.1, pp. 53–92.

— (2021). *Solar energetic particles: a modern primer on understanding sources, acceleration and propagation*. Springer.

Regnier, Stephane and Tahar Amari (2004). "3D magnetic configuration of the H filament and X-ray sigmoid in NOAA AR 8151". In: *Astronomy & Astrophysics* 425.1, pp. 345–352.

Riggi, S et al. (2025). "Solar flare forecasting with foundational transformer models across image, video, and time-series modalities". In: *Astronomy and Computing*, p. 101042.

Rishbeth, Henry and Owen K Garriott (1969). "Introduction to ionospheric physics". In: *Introduction to ionospheric physics*.

Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *The annals of mathematical statistics*, pp. 400–407.

Rombach, Robin et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer, pp. 234–241.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain." In: *Psychological review* 65.6, p. 386.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Rust, David M and Ashok Kumar (1996). "Evidence for helically kinked magnetic flux ropes in solar eruptions". In: *The Astrophysical Journal Letters* 464.2, pp. L199–L202.

Rust, DM and BJ LaBonte (2005). "Observational evidence of the kink instability in solar filament eruptions and sigmoids". In: *The Astrophysical Journal Letters* 622.1, pp. L69–L72.

Ryan, James M, John A Lockwood, and Hermann Debrunner (2000). "Solar energetic particles". In: *Space Science Reviews* 93.1, pp. 35–53.

Sakoe, Hiroaki and Seibi Chiba (1978). "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE transactions on acoustics, speech, and signal processing* 26.1, pp. 43–49.

Sakurai, Kunitomo (1970). "On the magnetic configuration of sunspot groups which produce solar proton flares". In: *Planetary and Space Science* 18.1, pp. 33–40.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller (2017). "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". In: *arXiv preprint arXiv:1708.08296*.

Sammis, Ian, Frances Tang, and Harold Zirin (2000). "The dependence of large flare occurrence on the magnetic structure of sunspots". In: *The Astrophysical Journal* 540.1, pp. 583–587.

Sato, Hiroatsu et al. (2019). "Solar radio burst events on 6 September 2017 and its impact on GNSS signal frequencies". In: *Space Weather* 17.6, pp. 816–826.

Savcheva, AS et al. (2012). "Photospheric flux cancellation and the build-up of sigmoidal flux ropes on the Sun". In: *The Astrophysical Journal* 759.2, p. 105.

Scherliess, Ludger (2016). "Ionospheric Response to X-Ray and EUV Flux Changes During Solar Flares: A Review". In: *Ionospheric Space Weather: Longitude and Hemispheric Dependences and Lower Atmosphere Forcing*, pp. 225–242.

Scherrer, Philip Hanby et al. (2012). "The helioseismic and magnetic imager (HMI) investigation for the solar dynamics observatory (SDO)". In: *Solar Physics* 275.1, pp. 207–227.

Schrijver, Carolus J (2007). "A characteristic magnetic field pattern associated with all major solar flares and its use in flare forecasting". In: *The Astrophysical Journal Letters* 655.2, pp. L117–L120.

— (2009). "Driving major solar flares and eruptions: A review". In: *Advances in Space Research* 43.5, pp. 739–755.

Schwenn, Rainer (2006). "Space weather: The solar perspective". In: *Living reviews in solar physics* 3.1, pp. 1–72.

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shorten, Connor and Taghi M Khoshgoftaar (2019). "A survey on image data augmentation for deep learning". In: *Journal of big data* 6.1, pp. 1–48.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034*.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sobel, Irwin, Gary Feldman, et al. (1968). "A 3x3 isotropic gradient operator for image processing". In: *a talk at the Stanford Artificial Project in* 1968, pp. 271–272.

Song, Hui et al. (2009). "Statistical assessment of photospheric magnetic features in imminent solar flare predictions". In: *Solar Physics* 254.1, pp. 101–125.

Sorriso-Valvo, Luca et al. (2015). "Sign singularity and flares in solar active region NOAA 11158". In: *The Astrophysical Journal* 801.1, p. 36.

Spelmen, Vimalraj S and R Porkodi (2018). "A review on handling imbalanced data". In: *2018 international conference on current trends towards converging technologies (ICCTCT)*. IEEE, pp. 1–11.

Sreeja, V, M Aquino, and Kees de Jong (2013). "Impact of the 24 September 2011 solar radio burst on the performance of GNSS receivers". In: *Space Weather* 11.5, pp. 306–312.

Srivastava, Nitish et al. (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Sterling, Alphonse C and Hugh S Hudson (1997). "Yohkoh SXT observations of X-ray "dimming" associated with a halo coronal mass ejection". In: *The Astrophysical Journal Letters* 491.1, pp. L55–L58.

Sterling, Alphonse C et al. (2000). "Yohkoh SXT and SOHO EIT observations of sigmoid-to-arcade evolution of structures associated with halo coronal mass ejections". In: *The Astrophysical Journal* 532.1, pp. 628–647.

Sudol, Jeffrey J and JW Harvey (2005). "Longitudinal magnetic field changes accompanying solar flares". In: *The Astrophysical Journal* 635.1, pp. 647–658.

Sun, Xudong (2013). "On the coordinate system of Space-Weather HMI Active Region Patches (SHARPs): a technical note". In: *arXiv preprint arXiv:1309.2392*.

Sun, Xudong and Mark CM Cheung (2021). "Non-Neutralized Electric Current of Active Regions Explained as a Projection Effect". In: *Solar Physics* 296.1, p. 7.

Sun, Xudong et al. (2012). "Evolution of magnetic field and energy in a major eruptive active region based on SDO/HMI observation". In: *The Astrophysical Journal* 748.2, p. 77.

Sun, Zeyu et al. (2022). "Predicting solar flares using CNN and LSTM on two solar cycles of active region data". In: *The Astrophysical Journal* 931.2, p. 163.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems* 27.

Syntelis, Petros et al. (2016). "The spectroscopic imprint of the pre-eruptive configuration resulting into two major coronal mass ejections". In: *Astronomy & Astrophysics* 588, A16.

Szegedy, Christian et al. (2015). "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tanaka, Katsuo (1991). "Studies on a very flare-active δ group: Peculiar δ spot evolution and inferred subsurface magnetic rope structure". In: *Solar Physics* 136.1, pp. 133–149.

Ternullo, M et al. (2006). "A statistical analysis of sunspot groups hosting M and X flares". In: *Astronomische Nachrichten: Astronomical Notes* 327.1, pp. 36–43.

Terven, Juan et al. (2025). "A comprehensive survey of loss functions and metrics in deep learning". In: *Artificial Intelligence Review* 58.7, p. 195.

The SunPy Community et al. (2020). "The SunPy Project: Open Source Development and Status of the Version 1.0 Core Package". In: *The Astrophysical Journal* 890.1. DOI: 10.3847/1538-4357/ab4f7a.

Toriumi, S et al. (2014). "Formation of a flare-productive active region: observation and numerical simulation of NOAA AR 11158". In: *Solar Physics* 289.9, pp. 3351–3369.

Toriumi, Shin and Haimin Wang (2019). "Flare-productive active regions". In: *Living Reviews in Solar Physics* 16.1, p. 3.

Toriumi, Shin et al. (2013). "The magnetic systems triggering the M6. 6 class solar flare in NOAA active region 11158". In: *The Astrophysical Journal* 773.2, p. 128.

Toriumi, Shin et al. (2017). "Magnetic properties of solar active regions that govern large solar flares and eruptions". In: *The Astrophysical Journal* 834.1, p. 56.

Tripathi, Durgesh et al. (2009). "Temperature tomography of a coronal sigmoid supporting the gradual formation of a flux rope". In: *The Astrophysical Journal* 698.1, pp. L27–L32.

Tsurutani, BT et al. (2005). "The October 28, 2003 extreme EUV solar flare and resultant extreme ionospheric effects: Comparison to other Halloween events and the Bastille Day event". In: *Geophysical Research Letters* 32.3.

Tsurutani, BT et al. (2009). "A brief review of "solar flare effects" on the ionosphere". In: *Radio Science* 44.01, pp. 1–14.

Van Den Oord, Aaron, Oriol Vinyals, et al. (2017). "Neural discrete representation learning". In: *Advances in neural information processing systems* 30.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Venkatakrishnan, P and Sanjiv Kumar Tiwari (2009). "On the absence of photospheric net currents in vector magnetograms of sunspots obtained from Hinode (solar optical telescope/spectro-polarimeter)". In: *The Astrophysical Journal* 706.1, pp. L114–L119.

Vincent, Pascal et al. (2008). "Extracting and composing robust features with denoising autoencoders". In: *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103.

Waldmeier, M (1938). "Chromosphärische Eruptionen. I. Mit 6 Abbildungen." In: *Zeitschrift für Astrophysik, Vol. 16, p. 276* 16, p. 276.

Walterscheid, RL (1989). "Solar cycle effects on the upper atmosphere-Implications for satellite drag". In: *Journal of spacecraft and rockets* 26.6, pp. 439–444.

Wang, Haimin (2006). "Rapid changes of photospheric magnetic fields around flaring magnetic neutral lines". In: *The Astrophysical Journal* 649.1, pp. 490–497.

Wang, Haimin and Chang Liu (2015). "Structure and evolution of magnetic fields associated with solar eruptions". In: *Research in Astronomy and Astrophysics* 15.2, pp. 145–174.

Wang, Haimin et al. (1991). "Motions, fields, and flares in the 1989 March active region". In: *Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 380, Oct. 10, 1991, p. 282-286. Research supported by NNSFC.* 380, pp. 282–286.

Wang, Haimin et al. (2002). "Core and large-scale structure of the 2000 November 24 X-class flare and coronal mass ejection". In: *The Astrophysical Journal* 569.2, pp. 1026–1031.

Wang, Haimin et al. (2008). "Study of magnetic channel structure in active region 10930". In: *The Astrophysical Journal* 687.1, pp. 658–667.

Wang, Haimin et al. (2018a). "Strong transverse photosphere magnetic fields and twist in light bridge dividing delta sunspot of active region 12673". In: *arXiv preprint arXiv:1801.02928*.

Wang, Jingxiu and Wei Li (1998). "Vector Magnetic Fields at the Base of Filaments and the Filament Environment". In: *International Astronomical Union Colloquium*. Vol. 167. Cambridge University Press, pp. 98–101.

Wang, Jingxiu et al. (1996). "Flares and the magnetic nonpotentiality". In: *Astrophysical Journal v. 456, p. 861* 456, p. 861.

Wang, Rui et al. (2015). "The role of active region coronal magnetic field in determining coronal mass ejection propagation direction". In: *The Astrophysical Journal* 814.1, p. 80.

Wang, Tongjiang, Aoao Xu, and Hongqi Zhang (1994). "Evolution of vector magnetic fields and vertical currents and their relationship with solar flares in AR 5747". In: *Solar Physics* 155.1, pp. 99–112.

Wang, Xiantong et al. (2020). "Predicting solar flares with machine learning: Investigating solar cycle dependence". In: *The Astrophysical Journal* 895.1, p. 3.

Wang, Xiaolong et al. (2018b). "Non-local neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803.

Wang, Yuming and Jie Zhang (2008). "A statistical study of solar active regions that produce extremely fast coronal mass ejections". In: *The Astrophysical Journal* 680.2, pp. 1516–1522.

Wang, Zhou and Alan C Bovik (2009). "Mean squared error: Love it or leave it? A new look at signal fidelity measures". In: *IEEE signal processing magazine* 26.1, pp. 98–117.

Wang, Zhou et al. (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4, pp. 600–612.

Warwick, Constance S (1966). "Sunspot configurations and proton flares". In: *Astrophysical Journal, vol. 145, p. 215* 145, p. 215.

Werbos, Paul J (2002). "Backpropagation through time: what it does and how to do it". In: *Proceedings of the IEEE* 78.10, pp. 1550–1560.

Wilkinson, Ladye K, AG Emslie, and GA Gary (1992). "On neutralized currents in the solar corona". In: *Astrophysical Journal, Part 2-Letters (ISSN 0004-637X), vol. 392, no. 1, June 10, 1992, p. L39-L42.* 392, pp. L39–L42.

Woodcock, Frank (1976). "The evaluation of yes/no forecasts for scientific and administrative purposes". In: *Monthly Weather Review* 104.10, pp. 1209–1214.

Woods, Thomas N et al. (2005). "Solar EUV Experiment (SEE): Mission overview and first results". In: *Journal of Geophysical Research: Space Physics* 110.A1.

Woods, Thomas N et al. (2012). "Extreme Ultraviolet Variability Experiment (EVE) on the Solar Dynamics Observatory (SDO): Overview of science objectives, instrument design, data products, and model developments". In: *The solar dynamics observatory*, pp. 115–143.

Xu, Ao-ao et al. (1991). "Statistical properties of solar gamma-ray flares". In: *Chinese astronomy and astrophysics* 15.3, pp. 318–328.

Xu, Jiyao et al. (2011). "The effect of periodic variations of thermospheric density on CHAMP and GRACE orbits". In: *Journal of Geophysical Research: Space Physics* 116.A2.

Xu, Kelvin et al. (2015). "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. PMLR, pp. 2048–2057.

Yasyukevich, Yu et al. (2018). "The 6 September 2017 X-class solar flares and their impacts on the ionosphere, GNSS, and HF radio wave propagation". In: *Space Weather* 16.8, pp. 1013–1027.

Yi, Kangwoo et al. (2021). "Visual explanation of a deep learning solar flare forecast model and its relationship to physical parameters". In: *The Astrophysical Journal* 910.1, p. 8.

Yurchyshyn, VB, VI Abramenko, and Vincenzo Carbone (2000). "Flare-related changes of an active region magnetic field". In: *The Astrophysical Journal* 538.2, pp. 968–979.

Zheng, Yanfang, Xuebao Li, and Xinshuo Wang (2019). "Solar flare prediction with the hybrid deep convolutional neural network". In: *The Astrophysical Journal* 885.1, p. 73.

Zirin, H and Haimin Wang (1993a). "Narrow lanes of transverse magnetic field in sunspots". In: *Nature* 363.6428, pp. 426–428.

Zirin, Harold and Margaret A Liggett (1987). "Delta spots and great flares". In: *Solar Physics* 113.1, pp. 267–283.

Zirin, Harold and Katsuo Tanaka (1973). "The flares of August 1972". In: *Solar Physics* 32.1, pp. 173–207.

Zirin, Harold and Haimin Wang (1993b). "Strong transverse fields in δ-spots". In: *Solar physics* 144.1, pp. 37–43.

Zolesi, Bruno and Ljiljana R Cander (2013). "Ionospheric Models for Prediction and Forecasting". In: *Ionospheric Prediction and Forecasting*. Springer, pp. 101–121.

Zwaard, Rens Van der et al. (2021). "Segmentation of coronal features to understand the solar EUV and UV irradiance variability III. Inclusion and Analysis of Bright points". In: *Solar Physics* 296.9, p. 138.