

Physical Interpretability of Deep Learning Models for Solar Flare Prediction

Maikel van den Broek

Chairperson: Jian Guo

Responsible Thesis Supervisor: Iklim Akay

Independent Examiner: Sijme-Jan Paardekooper

To be defended 28-08-2026

Abstract

Solar flares are sudden releases of energy from the Sun that can produce intense electromagnetic radiation and energetic particles, with the potential to disrupt satellite systems, radio communications, and other space-based and technological infrastructure. Accurate solar flare prediction is therefore important for providing timely warnings, while interpretable models are needed to understand the physical information driving predictions and to assess their reliability in operational settings. This study investigates a neural-network-based solar flare prediction model using SHARP magnetic-field parameters and flare-history information, with particular emphasis on model performance, interpretability, and the effects of incorporating physical constraints. The baseline model achieves competitive predictive performance, while temporal deployment on new data results in a small performance decrease. Physics-informed loss functions are shown to alter the model’s classification behavior without substantially reducing overall performance, although the effect depends strongly on the physical relationship imposed. The Pearson MEANGBZ constraint reduces false negatives at the cost of more false positives, whereas the Kendall MEANGAM constraint produces the opposite behavior and substantially more missed flares. These results demonstrate that TSS alone is insufficient to characterize the operational behavior of flare prediction models and that threshold selection and the balance between false positives and false negatives are important considerations. Finally, a parameter study shows substantial redundancy among the input features: a reduced model using only 11 features achieves a TSS of 0.7152, close to the approximately 0.72 of the full baseline. Overall, the results indicate that accurate, interpretable, and operational solar flare prediction can be achieved with fewer input parameters, while physics-informed learning can be used as a constraint to modify model behavior without necessarily compromising predictive performance.

Contents

List of Tables	4
List of Figures	6
List of Abbreviations	7
List of Symbols	8
1 Literature Review	16
1.1 Space weather and its consequences	16
1.1.1 Solar flares	16
1.1.2 The effect of solar flares	18
1.1.3 What causes solar flares?	19
1.2 Available data	20
1.2.1 Solar Dynamic Observatory	20
1.2.2 Helioseismic and Magnetic Imager	21
1.2.3 SHARP parameters	22
1.2.4 Geostationary Operational Environmental Satellite	24
1.3 Solar flare forecasting	24
1.3.1 Physics-based models	24
1.3.2 Statistical models	25
1.3.3 Machine learning models	25
1.3.4 Deep Learning	26
1.3.5 Overview	27
1.4 Interpretability in Deep Learning	29
1.4.1 Why interpretability matters.	29
1.4.2 Physics informed deep learning models	29
1.4.3 Physical interpretability in solar-flare prediction	30
1.5 Chapter overview	30
2 Research Proposal	32
2.1 Knowledge gap	32
2.2 Research objectives, questions, and hypotheses	32
2.3 Planning	34
2.4 Requirements	34
2.4.1 Overarching requirements	35
2.4.2 Dataset requirements	35
2.4.3 Physical-parameter requirements	35
2.4.4 Standard-model requirements	36
2.4.5 Physics-informed-model requirements	36
2.4.6 Computational requirements	36
2.5 Chapter overview	37

3	Dataset Construction	38
3.1	Magnetic field paramters	38
3.2	History paramters	40
3.3	Solar phase constant	40
3.4	Normalization	41
3.5	Label construction	41
3.6	Dataset Split	42
3.7	Chapter overview	42
4	Methodology	43
4.1	Model design	43
4.2	Physical loss function motivation	47
4.3	Evaluation metrics	47
4.4	Hyperparameter search	48
4.5	Feature study	49
4.6	Chapter overview	50
5	Results	52
5.1	Hyperparameter tuning	52
5.2	Baseline model	61
5.3	Physics-informed model	67
5.3.1	Best model case study: Pearson MEANGBZ	68
5.3.2	Worst model case study: Kendall MEANGAM	73
5.4	Parameter study	78
6	Discussion	82
6.1	Hyperparameter selection	82
6.2	Baseline model	83
6.3	Physics-informed model	85
6.4	Parameter study	89
6.5	Chapter overview	93
7	Verification	94
7.1	Hyperparameter verification	94
7.2	Verification using 2026 data	94
7.3	Probability distribution verification	94
7.4	Verification of the 11-feature model	95
8	Conclusion	96
8.1	Future outlook	99
9	Appendix	107

List of Tables

1	Solar Flare Classes	17
2	Solar flares and Radio Blackouts	18
3	SHARP parameters	23
4	Qualitative comparison of solar flare forecasting model categories.	28
5	Selected hyperparameters from Optuna tuning and their corresponding values used in final training.	60
6	Confusion Matrix baseline model on test set	65
7	Confusion Matrix baseline model 2026 data using threshold 0.097	67
8	Δ TSS (+ better) for different correlation measures.	68
9	Confusion Matrix Pearson MEANGBZ model with threshold 0.1820	68
10	Confusion Matrix Kendall MEANGAM model at threshold 0.2173	73
11	Confusion Matrix baseline model with only 11 features at threshold 0.2173	78
12	Overview of comparable models also predicting M and X in a 24 window.	83

List of Figures

1	Solar Flare in 5 Wavelengths	17
2	Solar Cycle overview	20
3	SDO spacecraft	21
4	The Zeeman Effect	22
5	LSTM model architecture overview [75]	43
6	Mean TSS per completed trial	52
7	Hidden size TSS scatter plot	53
8	Number of layers TSS scatter plot	54
9	Dropout rate TSS scatter plot	55
10	Learning rate TSS scatter plot	56
11	Weight decay TSS scatter plot	57
12	Positive weight TSS scatter plot	58
13	TSS score distribution per fold	59
14	Loss during training of the baseline model	61
15	Validation TSS of baseline model during training	62
16	Threshold evolution of baseline model during training	63
17	Class distribution	64
18	TSS per threshold	65
19	F1 per threshold	66
20	Loss during training of the Pearson MEANGBZ model	69
21	Physics informed loss during training of the Pearson MEANGBZ model	70
22	Validation TSS of Pearson MEANGBZ during training	71
23	Class distribution Pearson	72
24	Loss during training of the Kendall MEANGAM model	74
25	Validation TSS of the Kendall MEANGAM model during training	75
26	Physics informed loss during training of the Kendall MEANGAM model	76
27	Class distribution Kendall	77
28	Heatmap first 13 steps parameter elimination	79
29	Heatmap last 13 steps parameter elimination	80
30	Model trained with 11 parameters	81
31	Physics correlation baseline model	108
32	Saliency baseline history	109
33	Saliency baseline delta	110
34	Saliency baseline SHARP	111
35	Saliency Pearson history	112
36	Saliency Pearson delta	113
37	Saliency Pearson SHARP	114
38	Physics correlation Pearson MEANGBZ model	115
39	Saliency Kendall history	116
40	Saliency Kendall delta	117
41	Saliency Kendall SHARP	118
42	Physics correlation Kendall MEANGAM model	119
43	TSS first 13 steps parameter elimination	120
44	TSS last 13 steps parameter elimination	121

45	TSS permutation per eliminations step	122
----	---	-----

List of Abbreviations

CME	Coronal Mass Ejection
NASA	National Aeronautics and Space Administration
GOES	Geostationary Operational Environmental Satellite
NOAA	National Oceanic and Atmospheric Administration
EUV	Extreme Ultra Violet
HF	High Frequency
UV	Ultra Violet
SDO	Solar Dynamic Observatory
AIA	Atmospheric Imaging Assembly
EVE	Extreme Ultraviolet Variability Experiment
HMI	Helioseismic and Magnetic Imager
SHARP	Space Weather HMI Active Region Patches
EXIS	Extreme Ultraviolet and X-ray Irradiance Sensors
MHD	Magnetohydrodynamics
SOC	Self-organized Criticality
SOHO	Solar and Heliospheric Observatory
HDI	Michelson Doppler Imager
NN	Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
PINN	Physics Informed Neural Network
NRT	Near Real Time
CEA	Cylindrical Equal-Area
HEK	Heliophysics Event Knowledgebase
TSS	True Skill Statistic
LOWESS	Locally Weighted Scatterplot Smoothing

List of Symbols

Magnetic field quantities

- Φ Total unsigned magnetic flux
- B_x x -component of the magnetic field
- B_y y -component of the magnetic field
- B_z Vertical component of the magnetic field
- B_h Horizontal component of the magnetic field
- B_{tot} Total magnetic field
- B^{Obs} Observed magnetic field
- B^{Pot} Potential magnetic field
- γ Magnetic field inclination angle
- $\bar{\gamma}$ Mean magnetic field inclination angle
- ∇B_{tot} Gradient of the total magnetic field
- ∇B_z Gradient of the vertical magnetic field
- ∇B_h Gradient of the horizontal magnetic field
- ∂_x Partial derivative with respect to the x -coordinate
- ∂_y Partial derivative with respect to the y -coordinate

Electric current, helicity and magnetic energy

- J_z Vertical electric current density
- $J_{z,\text{tot}}$ Total unsigned vertical electric current
- J_{sum} Net electric current per magnetic polarity
- α Magnetic twist or force-free parameter
- H_c Current helicity density
- $\overline{H_c}$ Mean current helicity density
- $H_{c,\text{tot}}$ Total unsigned current helicity
- $H_{c,\text{abs}}$ Absolute net current helicity
- $\bar{\rho}$ Mean excess magnetic energy density
- ρ_{tot} Total excess magnetic energy

Γ Magnetic shear angle

$\bar{\Gamma}$ Mean magnetic shear angle

A_{AR} Area of the active region

Spatial quantities and mathematical operators

dA Differential area element

N Number of pixels or measurements used in an average

$|\cdot|$ Absolute value or magnitude

$\sum$ Summation operator

$\sqrt{\cdot}$ Square-root operation

Δ Difference operator

∇ Spatial gradient operator

$\arctan$ Inverse tangent function

$\arccos$ Inverse cosine function

$\propto$ Proportional to

$\approx$ Approximately equal to

$\in$ Element-of relation

$\cup$ Set union

mod Modulo operation

π Mathematical constant pi

e Base of the natural logarithm

$\log$ Natural logarithm

Units

Mx Maxwell, unit of magnetic flux

G Gauss, unit of magnetic-field strength

Mm Megametre

mA, m^{-2} Milliampere per square metre

A Ampere

G^2m^{-1} Gauss squared per metre

erg Unit of energy in the CGS system
 erg cm^{-3} Energy density in the CGS system
 μHem Microhemisphere, a unit of projected solar area
pixels Individual image elements in the magnetogram

Data and indexing notation

r Active-region index
 t Time point or time index
 t_0 Earlier time point used for interpolation
 t_1 Later time point used for interpolation
 t_{ref} Reference time corresponding to the most recent known solar minimum
 t_i Time point associated with sample or observation i
 t_k Time at which a flare occurred
 t' Second time-step index used in pairwise comparisons
 i Sample or observation index
 j Feature or parameter index
 x Generic feature
 x_t Feature value at time t
 $x^{(j)}$ Feature j
 $x_{r,t}^{(j)}$ Feature j for active region r at time t
 $\hat{x}_{r,t}^{(j)}$ Interpolated value of feature j for active region r at time t
 $\Delta x_{r,t}$ Change in feature x for active region r between consecutive time steps
 $X_{i,t,j}$ Value of feature j for sample i at time t
 $\tilde{X}_{i,t,j}$ Normalized value of feature j for sample i at time t
 V_{train} Set of non-padding observations in the training set
 $|V_{\text{train}}|$ Number of non-padding observations in the training set

Flare history notation

F Set of all previous flares in an active region
 F_B Set of previous B-class flares

F_C Set of previous C-class flares
 F_M Set of previous M-class flares
 F_X Set of previous X-class flares
 f_i Individual solar flare
 f_j Individual solar flare indexed by j
 $E(f_i)$ Energy or magnitude of flare f_i
 τ Flare-history decay constant
 $\text{Bdec}(x_t)$ Decay-weighted history of B-class flares
 $\text{Cdec}(x_t)$ Decay-weighted history of C-class flares
 $\text{Mdec}(x_t)$ Decay-weighted history of M-class flares
 $\text{Xdec}(x_t)$ Decay-weighted history of X-class flares
 $\text{Edec}(x_t)$ Decay-weighted total flare-energy feature
 $\text{logEdec}(x_t)$ Logarithmic decay-weighted total flare-energy feature

Solar-cycle notation

$P_\odot$ Solar-cycle period
 $\phi(t)$ Solar-cycle phase at time t
 $\sin$ Sine function
 $\cos$ Cosine function
 2π Full angular cycle in radians
 $[-1, 1]^2$ Two-dimensional range of the sine/cosine phase encoding

Normalization notation

μ_j Mean of feature j calculated from the training set
 σ_j Standard deviation of feature j calculated from the training set
 ϵ Small numerical constant used to prevent division by zero
 10^{-8} Value assigned to ϵ during normalization

Model architecture and sequence notation

B Batch size
 T Sequence length or number of time steps

P Number of input features
 e Training epoch or step index
 x_t Input feature vector at time step t
 y_i Target label for sample i
 $\hat{y}$ Model output logit
 p Predicted probability
 h_t LSTM hidden state at time step t
 h_{t-1} LSTM hidden state at the previous time step
 h_T Final LSTM hidden state
 H Sequence of hidden states produced by the LSTM
 c_t LSTM cell state at time step t
 c_{t-1} LSTM cell state at the previous time step
 $\tilde{c}_t$ Candidate LSTM cell state at time step t
LSTM gates and parameters
 f_t Forget-gate activation at time step t
 i_t Input-gate activation at time step t
 o_t Output-gate activation at time step t
 W_f Forget-gate weight matrix
 W_i Input-gate weight matrix
 W_c Candidate cell-state weight matrix
 W_o Output-gate weight matrix
 b_f Forget-gate bias
 b_i Input-gate bias
 b_c Candidate cell-state bias
 b_o Output-gate bias
 W_1 First feedforward-layer weight matrix
 W_2 Second feedforward-layer weight matrix
 b_1 First feedforward-layer bias

b_2 Second feedforward-layer bias

Physics-based head and correlation notation

s_t Physics-head logit at time step t

$\hat{p}_t$ Physics-head confidence at time step t

c_t^j Cumulative value of selected physical feature j up to time t

W_s Physics-head weight matrix

b_s Physics-head bias

u, v Generic variables used for correlation calculations

$\bar{u}$ Mean of variable u

$\bar{v}$ Mean of variable v

$\tilde{r}_t(u)$ Soft rank of u_t

r_P Pearson correlation coefficient

$\hat{\tau}$ Differentiable approximation of Kendall's correlation coefficient

τ_{rank} Temperature parameter used for differentiable ranking

Loss functions

$\mathcal{L}$ Total training loss

$\mathcal{L}_{\text{BCE}}$ Weighted binary cross-entropy loss

$\mathcal{L}_{\text{phys}}$ Physics-based loss

$\mathcal{L}_{\text{phys}}^{(i)}$ Physics-based loss for sample i

α Weight controlling the relative contribution of the physics-based loss

w^+ Positive-class weight in the binary cross-entropy loss

Corr Correlation measure used in the physics-based loss

Activation functions and neural-network operations

$\sigma(\cdot)$ Sigmoid activation function

$\tanh(\cdot)$ Hyperbolic tangent activation function

$\text{ReLU}(\cdot)$ Rectified linear unit activation function

$\text{LN}(\cdot)$ Layer-normalization operation

$\text{Dropout}(\cdot)$ Dropout operation

$\odot$ Element-wise multiplication

$|\cdot|_2$ Euclidean (L_2) norm

Optimizer notation

θ Model parameters

$\theta^{(e)}$ Model parameters at training step or epoch e

$\theta^{(e+1)}$ Updated model parameters

η Learning rate

λ Weight-decay coefficient

$\hat{\mathbf{m}}$ Bias-corrected first moment estimate in AdamW

$\hat{\mathbf{v}}$ Bias-corrected second moment estimate in AdamW

ϵ_{Adam} Small numerical constant used for numerical stability in AdamW

∇ Gradient operator

Classification and evaluation metrics

TP True positives

FP False positives

TN True negatives

FN False negatives

Precision Fraction of predicted positives that are true positives

Recall Fraction of actual positives correctly predicted

TSS True Skill Statistic

Physical model notation

$\mathbf{B}$ Magnetic field vector

$\mathbf{J}$ Electric current density vector

$\mathbf{E}$ Electric field vector

μ_0 Vacuum permeability

ϵ_0 Vacuum permittivity

$\nabla \times \mathbf{B}$ Curl of the magnetic field

$\frac{\partial \mathbf{E}}{\partial t}$ Partial time derivative of the electric field

$sign(x)$ Sign function

Dataset and experimental constants

0 Negative class label or padding value, depending on context

1 Positive class label

24 Length of the input and prediction windows in hours

60% Fraction of the dataset assigned to training

20% Fraction of the dataset assigned to validation

20% Fraction of the dataset assigned to testing

70° Maximum allowed angular distance from the solar-disk centre

5 Maximum allowed interpolation gap in time steps

12 Flare-history decay constant τ

11 Approximate length of the solar cycle in years

4015 Approximate solar-cycle period in days

1 Literature Review

1.1 Space weather and its consequences

Today’s civilization is increasingly dependent on space-based infrastructure. Satellite communications, Global Navigation Satellite Systems, weather forecasting, Earth observation, financial timing networks, aviation and electrical power grids are all susceptible to disturbances caused by space weather.

The economic consequences of severe space weather are substantial. Estimates suggest that an event comparable to the 1859 Carrington Event, the most intense geomagnetic storm in recorded history, could result in economic losses ranging from hundreds of billions to several trillion U.S. dollars, depending on the affected infrastructure and geographic scope, with recovery potentially taking several years [1]. Even moderate events can result in significant operational costs due to satellite anomalies, increased atmospheric drag, and service interruptions.

A recent example demonstrating these impacts occurred in February of 2022, when a geomagnetic storm caused the premature loss of 38 out of 49 Starlink satellites shortly after launch. The storm heated and expanded Earth’s upper atmosphere, dramatically increasing atmospheric drag at the satellites’ low insertion orbit. Because the satellites were unable to raise their altitude in time, they re-entered the atmosphere and burned up, resulting in a loss estimated at over 50 million USD [2]. This shows the vulnerability of modern satellite constellations to space weather.

From the perspective of space engineering, understanding and forecasting solar activity is therefore essential. Spacecraft must be designed to withstand enhanced radiation, charging effects, and increased drag, while satellite operators require accurate forecasts to implement mitigation strategies such as delaying launches, modifying spacecraft operations, or protecting sensitive electronics. As humanity’s reliance on space infrastructure continues to grow, improving the ability to predict and understand solar eruptions becomes even more important.

The principal drivers of adverse space weather are solar flares and coronal mass ejections (CME). Although often associated with one another, they are distinct manifestations of the rapid release of magnetic energy stored in the Sun’s atmosphere. Solar flares produce intense bursts of electromagnetic radiation spanning the entire spectrum from radio waves to gamma rays, reaching Earth in approximately 8 minutes. In contrast, CMEs eject billions of tons of magnetized plasma into interplanetary space, typically arriving at Earth after 1–4 days, where they can interact with Earth’s magnetosphere and trigger geomagnetic storms. Although both phenomena can have severe consequences, the short travel time of solar flare radiation leaves little opportunity for mitigation, making accurate and timely forecasting essential.

1.1.1 Solar flares

Solar flares are large eruptions of electromagnetic radiation from the Sun lasting from minutes to hours [3]. Flares occur in active regions and are often, but not always, accompanied by CME’s, solar particle events, and other eruptive solar phenomena. The exact link to CME’s is currently not well understood [4]. An example of a flare emerging from an active region is displayed in Figure

1 from the National Aeronautics and Space Administration (NASA). As can be seen in the visible light photo, flares generally originate from active regions also called sunspots.

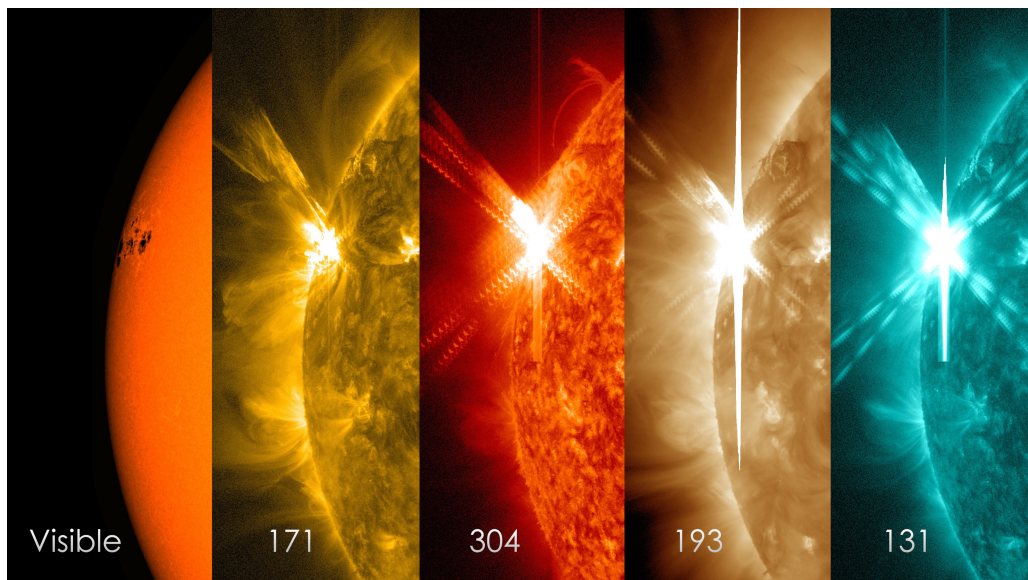


Figure 1: NASA’s Solar Dynamics Observatory captured these images of a solar flare, as seen in the bright flash on the left, on May 5, 2015. Each image shows a different wavelength of extreme ultraviolet light that highlights a different temperature of material on the sun. From left to right, the wavelengths are: visible light, 171 angstroms, 304 angstroms, 193 angstroms and 131 angstroms. Each wavelength has been colorized [5].

The electromagnetic radiation from flares ranges across the electromagnetic spectrum, from radio waves to gamma rays [6]. Solar flares are detected using various satellites and ground stations, most importantly the Geostationary Operational Environmental Satellite (GOES). This is a system of satellites that supports weather forecasting, severe storm tracking, and meteorology research. It is operated by the United States’ National Oceanic and Atmospheric Administration (NOAA)’s National Environmental Satellite, Data, and Information Service division. While flares can be detected using various wavelengths, the current state-of-the-art makes use of soft X-rays, with wavelengths of 0.1 to 0.8 nanometers, to classify solar flare intensity. The peak flux of soft X-rays determines the class of a solar flare, displayed in Table 1.

Table 1: Solar flare classification by peak flux.	
Classification	Peak flux range (W/m^2)
A	$< 10^{-7}$
B	$10^{-7} - 10^{-6}$
C	$10^{-6} - 10^{-5}$
M	$10^{-5} - 10^{-4}$
X	$\geq 10^{-4}$

Besides the class, a number denoting the strength within the class is added. This means an X2 flare is twice the strength of an X1 flare, an X3 flare is three times as powerful as an X1, and a C10 flare is noted as an M1 flare. In addition, any flare exceeding X10 will still be labeled within the X class. The strongest known flare is the Carrington event, which is estimated to be an X45 class flare [7]. The frequency distributions of multiple solar flare characteristics, such as peak fluxes of radio, extreme ultraviolet, and hard and soft X-ray emissions; total energies; and flare durations have all been found to follow a power law distribution [8, 9, 10].

1.1.2 The effect of solar flares

The electromagnetic radiation emitted during a solar flare propagates away from the Sun at the speed of light. As the radiation spreads through space, its irradiance decreases approximately with the inverse square of the distance from the source. The ionizing radiation, mainly the X-ray and extreme ultraviolet radiation (EUV), is known to affect planetary atmospheres and humans or spacecrafts in space. The increased level of X-ray and EUV radiation results in ionization in the lower layers of the ionosphere on the sunlit side of Earth.

Under normal conditions, high frequency (HF) radio waves can support communication over long distances by refraction via the upper layers of the ionosphere. When a strong enough solar flare occurs, ionization is produced in the lower, more dense layers of the ionosphere (the D-layer). Radio waves propagating through this layer experience increased absorption because frequent collisions between free electrons and neutral particles convert part of the wave energy into heat. This can cause HF radio signals to become degraded or completely absorbed. This results in a radio blackout, which is the absence of HF communication, primarily impacting the 3 to 30 MHz band [3]. Radio blackouts are classified using a five-level NOAA Space Weather Scale, directly related to the flare’s max peak in soft X-rays reached or expected [11]. This relation is displayed in Table 2

Table 2: Solar flare radio blackout classifications and their consequences [12]

Classification	Description	Associated	Consequences
R1	Minor radio blackout	M1	Weak signals may fade briefly; minor disruption to HF communications
R2	Moderate radio blackout	M5	HF radio communications disrupted for tens of minutes; some navigation systems may be affected
R3	Strong radio blackout	X1	Widespread HF communication loss; navigation and timing signals impacted
R4	Severe radio blackout	X10	Long-lasting HF communication blackout; critical systems such as aviation and emergency services affected
R5	Extreme radio blackout	X20	Complete HF radio failure; severe disruption to navigation, communication, and emergency operations

While enhanced ionization in the D-layer affects radio propagation, the enhanced EUV and X-ray flux also deposits energy higher in the atmosphere, heating the thermosphere. This heating

causes the atmosphere expand and increase the neutral atmospheric density at typical low Earth orbit (LEO) altitudes (approximately 200–800 km). The resulting increase in atmospheric density produces greater aerodynamic drag on satellites, accelerating orbital decay and increasing the frequency of orbit maintenance maneuvers. During periods of intense solar activity, the enhanced drag can substantially reduce satellite lifetimes, degrade orbit prediction accuracy, and complicate space traffic management. Consequently, thermospheric density variations driven by solar activity are a major consideration for satellite mission planning and operation [13, 14, 15].

In addition, large solar flares are frequently associated with CMEs, although either phenomenon can occur independently. When directed toward Earth, CMEs can interact with the magnetosphere and trigger geomagnetic storms. The most prominent example in recent history is the March 1989 magnetic storm, which had widespread effects on power systems, including a blackout of the Hydro-Québec system [16]. This storm was caused by two CMEs, the first associated with an X4.5 flare on 10 March and the second linked to an M7.3 flare on 12 March.

1.1.3 What causes solar flares?

Solar flares are among the most energetic phenomena in the Solar System, releasing up to 10^{25} J of energy over timescales ranging from minutes to several hours [17]. Although the precise physical mechanisms governing flare initiation and energy release remain an active area of research, there is broad consensus regarding the origin of the energy that powers these events. The energy released during a flare is believed to originate from the Sun’s magnetic field, specifically from magnetic energy that has accumulated in the solar corona due to the continuous motion of plasma in the photosphere [18, 19, 20].

The Sun’s differential rotation, together with convective motions beneath the photosphere, continually twists, stretches, and shears magnetic field lines emerging from the solar surface. These processes increase the amount of magnetic free energy stored in active regions, particularly around sunspots where magnetic fields are strongest [21]. As the magnetic stress increases, a thin current sheet may form between oppositely directed magnetic field lines. Within this narrow region, the magnetic field changes direction rapidly and magnetic reconnection can be initiated, allowing the stored magnetic energy to be released [22, 23].

The most widely accepted mechanism for the rapid release of this stored energy is magnetic reconnection [24, 25]. During magnetic reconnection, oppositely directed magnetic field lines are brought together, break, and reconnect into a lower-energy configuration. This topological rearrangement converts magnetic energy into kinetic energy, thermal energy, and the acceleration of charged particles. Although magnetic reconnection is generally accepted as the fundamental energy release mechanism, the detailed microphysical processes controlling the reconnection rate, particle acceleration, and energy partitioning remain not fully understood and continue to be investigated [4, 26].

The accelerated charged particles, primarily electrons and ions, propagate along magnetic field lines into the denser layers of the solar atmosphere, including the chromosphere. As these high-energy particles collide with the surrounding plasma, they deposit energy through Coulomb collisions, rapidly heating the local plasma to temperatures exceeding 10^7 K. This heating produces intense electromagnetic radiation across the spectrum, including radio waves, visible light, ultraviolet (UV)

, extreme ultraviolet (EUV), X-rays, and gamma rays [27]. The enhanced X-ray and EUV emissions are particularly important because they can significantly affect Earth's ionosphere and disrupt satellite communications, navigation systems, and radio transmissions.

Observations from both ground-based and space-based solar observatories provide strong evidence linking solar flares to regions of intense magnetic activity. Nearly all major flares originate in magnetically complex active regions associated with sunspots, where strong magnetic gradients and highly sheared magnetic field configurations are common [28]. This relationship explains the strong statistical correlation between flare occurrence and measures of magnetic complexity and field strength.

Solar flare activity also follows the approximately 11-year solar cycle, during which the Sun's global magnetic field periodically reverses polarity. During solar maximum, the number of active regions and sunspots increases substantially, resulting in a much higher frequency of solar flares. Conversely, during solar minimum, magnetic activity is greatly reduced, and flares occur much less frequently. This cyclic variation is illustrated in Figure 2.

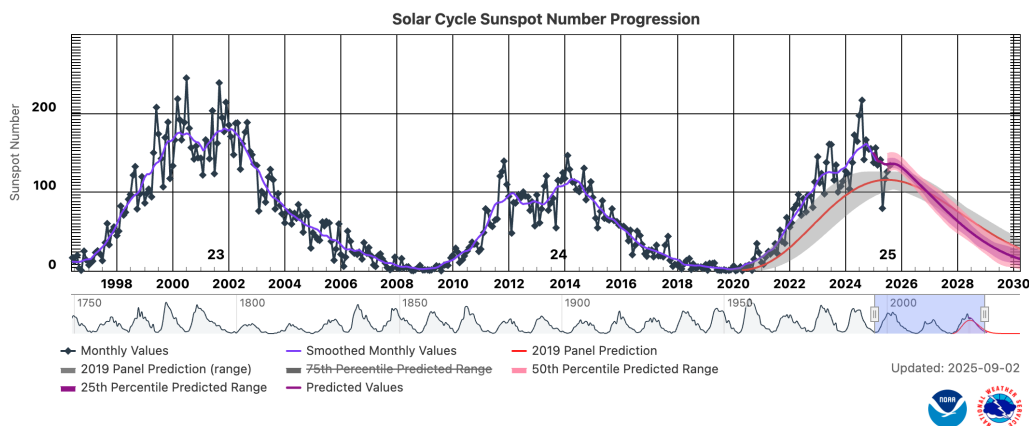


Figure 2: Amount of sunspots between 1996 and 2025. Depicting the cyclical behavior of solar activity. Image from the NOAA[29].

1.2 Available data

1.2.1 Solar Dynamic Observatory

The Solar Dynamics Observatory (SDO) was launched on 11 February 2010 aboard an Atlas V 401 launch vehicle [30]. Its primary objective is to improve the understanding of solar variability and space weather processes affecting Earth. Although originally designed for a five-year mission, it is expected to remain operational until approximately 2030.

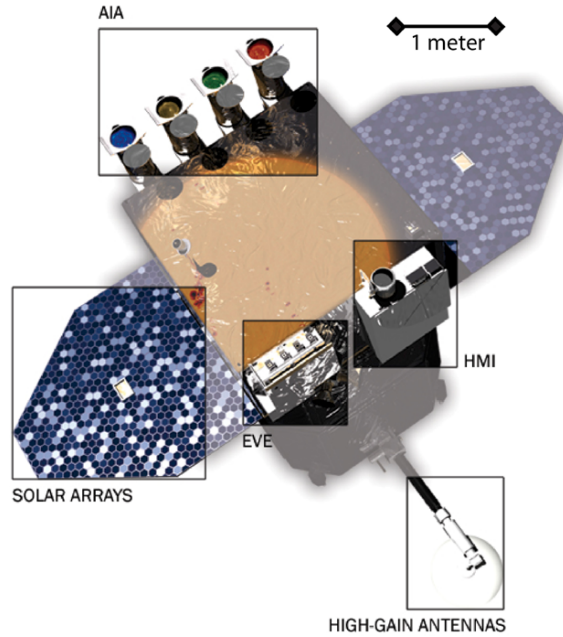


Figure 3: The SDO spacecraft with the instruments, high-gain antennas, and solar arrays highlighted. The thrusters and main engine are located on the opposite face of the satellite [30].

The SDO features 3 sets of instruments, each built by a different Science Investigation Team. These teams are the Atmospheric Imaging Assembly (AIA), the Extreme Ultraviolet Variability Experiment (EVE), and the Helioseismic and Magnetic Imager (HMI). Each group focuses on different aspects of the sun and its behavior.

1.2.2 Helioseismic and Magnetic Imager

The HMI investigation has three primary objectives: First, determine how and why the Sun varies; second, improve the understanding of how the Sun drives global change and space weather; and third, determine to what extent predictions of the space weather and global change can be made to prototype predictive techniques [31]. It is this third objective that represents the usefulness of this data for this particular research.

The HMI makes use of the Zeeman effect to measure the magnetic field of the sun. The Zeeman effect describes how atomic energy levels, and therefore the spectral lines, are modified in the presence of a magnetic field. In the absence of a magnetic field, an atomic transition produces a single spectral line at a well-defined wavelength. When a magnetic field is present, the atomic energy level is split into multiple sublevels, depicted in Figure 4. As a result, the spectral line divides into several closely spaced components.

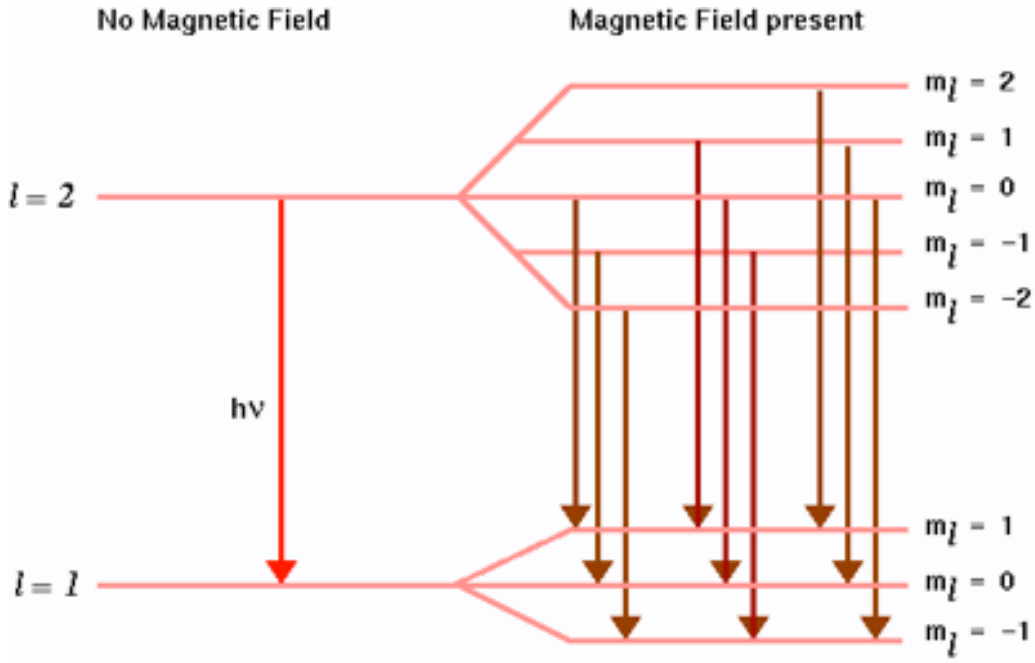


Figure 4: The Zeeman Effect. In the presence of a magnetic field, atomic energy levels get split [32]

The HMI makes use of the Fe I 6173 Å line [33]. This line splits into 3 components: a π component that is unshifted and two σ components that are oppositely shifted and circularly polarized. This polarization is described using the Stokes parameters I , V , Q and U . Where I encodes the intensity. V corresponds to circular polarization, which relates to the line of sight field (LoS). Q and U correspond to linear polarization (transverse field). The magnetic field parameters are reconstructed through an algorithm known as Very Fast Inversion of the Stokes Vector (VFISV), developed by Borrero et al. [34].

The data products of the HMI come in two sets produced by two data pipelines. One computes the LoS observables on a 45-second cadence. Multiple measurements are required to accurately measure the Q and U because of the low intensity and the background noise. So, the second pipeline computes time-averaged vector and LoS observables every 720 seconds. Both pipelines provide near-real-time (NRT) data products. Fully processed versions become available after additional calibration and processing. Observables are computed from sets of filtergrams taken in a fixed, repeating sequence called a frame list. The frame list specifies the wavelength and polarization state for each exposure [35].

1.2.3 SHARP parameters

Although HMI continuously observes the entire solar disk, full-disk data are not required when the objective is to study or predict solar flares. Since flares originate in magnetically active regions, the analysis can be restricted to these localized areas without losing the most relevant magnetic

information. The Space-weather HMI Active Region Patches (SHARP) data product automatically identifies active regions and extracts both the corresponding image data and a set of derived magnetic parameters, providing a more compact and computationally efficient representation than full-disk observations by substantially reducing the dimensionality of the data. However, this reduction may come at the cost of some information loss, as the derived parameters cannot fully capture the detailed spatial structure and complexity present in the original magnetograms. The SHARP data series includes the photospheric vector magnetic field and its uncertainty, along with Doppler velocity, continuum intensity, and line-of-sight magnetic field. In addition, SHARP provides a set of summary parameters, or keywords, derived for the identified active regions on the solar disk. These parameters are computed from the vector magnetograms, typically after remapping the data to a Cylindrical Equal Area (CEA) projection [36]. Several of these parameters have been found to exhibit a high statistical correlation with solar flares [37, 38]. Below is a brief description of each SHARP parameter used in this study. For brevity, proportionality constants are omitted where only the qualitative definition is relevant.

Table 3: Overview of SHARP summary parameters

SHARP Parameter	Meaning	Units	Formula
USFLUX	Total unsigned magnetic flux	Mx	$\Phi = \sum B_z dA$
MEANGAM	Mean inclination angle	deg	$\bar{\gamma} = \frac{1}{N} \sum \arctan\left(\frac{B_h}{B_z}\right)$
MEANGBT	Mean total field gradient	G Mm ⁻¹	$ \nabla B_{\text{tot}} = \frac{1}{N} \sum \sqrt{(\partial_x B_{\text{tot}})^2 + (\partial_y B_{\text{tot}})^2}$
MEANGBZ	Mean vertical field gradient	G Mm ⁻¹	$ \nabla B_z = \frac{1}{N} \sum \sqrt{(\partial_x B_z)^2 + (\partial_y B_z)^2}$
MEANGBH	Mean horizontal field gradient	G Mm ⁻¹	$ \nabla B_h = \frac{1}{N} \sum \sqrt{(\partial_x B_h)^2 + (\partial_y B_h)^2}$
MEANJZD	Mean vertical current density	mA m ⁻²	$\bar{J}_z \propto \frac{1}{N} \sum (\partial_x B_y - \partial_y B_x)$
TOTUSJZ	Total unsigned vertical current	A	$J_{z,\text{tot}} = \sum J_z dA$
MEANALP	Mean twist parameter	Mm ⁻¹	$\alpha \propto \sum J_z B_z \sum B_z^2$
MEANJZH	Mean current helicity density	G ² m ⁻¹	$\bar{H}_c \propto \frac{1}{N} \sum B_z J_z$
TOTUSJH	Total unsigned current helicity	G ² m ⁻¹	$H_{c,\text{tot}} \propto \sum B_z J_z $
ABSNJZH	Absolute net current helicity	G ² m ⁻¹	$H_{c,\text{abs}} \propto \left \sum B_z J_z \right $
SAVNCPP	Net current per polarity	A	$J_{\text{sum}} \propto \left \sum_{B_z > 0} J_z dA \right + \left \sum_{B_z < 0} J_z dA \right $
MEANPOT	Mean excess magnetic energy density	erg cm ⁻³	$\bar{\rho} \propto \frac{1}{N} \sum (B^{\text{Obs}} - B^{\text{Pot}})^2$
TOTPOT	Total excess magnetic energy	erg	$\rho_{\text{tot}} \propto \sum (B^{\text{Obs}} - B^{\text{Pot}})^2 dA$
MEANSHR	Mean shear angle	deg	$\bar{\Gamma} = \frac{1}{N} \sum \arccos\left(\frac{\mathbf{B}^{\text{Obs}} \cdot \mathbf{B}^{\text{Pot}}}{ \mathbf{B}^{\text{Obs}} \mathbf{B}^{\text{Pot}} }\right)$
SHRGT45	Area with shear > 45°	%	area with shear > 45° / HARP area
AREA_ACR	Area of the active region after correction	Mm ²	$A_{\text{AR}} = \sum dA$
SIZE_ACR	Area of the projected active region	μHem	Sum of projected pixel areas within the HARP mask
NPIX	Number of pixels in the active region mask	pixels	Sum of total pixels

1.2.4 Geostationary Operational Environmental Satellite

The Geostationary Operational Environmental Satellite, or GOES system, is a series of geostationary satellites operated by the United States' National Oceanic and Atmospheric Administration (NOAA)'s National Environmental Satellite, Data, and Information Service division. These satellites orbit Earth at an altitude of roughly 35,786 km, allowing them to remain in a fixed position relative to Earth's surface. This allows consistent monitoring of atmospheric conditions, weather systems, and space weather phenomena across the Western Hemisphere. It has been operational since 1975 and continues to provide valuable data.

The primary goal of the GOES program is to provide continuous observations of Earth's atmosphere and near-space environment to support weather forecasting and environmental research. The GOES satellites provide atmospheric measurements at high frequency, allowing for meteorologists to track storms, hurricanes, and severe weather in near real time.

In addition, the GOES system is used to detect and measure solar flares. For this, it makes use of the Extreme Ultraviolet and X-ray Irradiance Sensors (EXIS). Where the X-ray sensors conduct measurements of two wavelength bands (0.05–0.4 nm and 0.1–0.8 nm) [39]. This is the reason why solar flares use the 0.1–0.8 nm X-ray range for classification [40]. Because of its long-standing record, it continues to be the main metric used for solar flare detection and classification.

1.3 Solar flare forecasting

Short-term solar flare forecasting models typically provide predictions within 1 to 3 days before the occurrence of a flare. Generally, the forecast includes predictions such as if a flare will occur, the intensity of the flare, the time and location of its occurrence, and whether it will be associated with other solar activities. Solar flare prediction can be formulated either as a classification task or as a regression task. In classification, the model puts out discrete values such as "flare" or "no flare." In regression, continuous outcomes are generated, like the peak soft X-ray flux or the total flare duration [41].

1.3.1 Physics-based models

Physics-based models attempt to predict and simulate the behavior of the sun through the use of fundamental physical laws. There are two general types of models that are used, magnetohydrodynamics (MHD) and self-organized criticality (SOC).

MHD enables the modeling of interactions between solar plasma and electromagnetic fields across both temporal and spatial scales. While MHD models show success in variation of problems, the understanding of the underlying physics behind solar flares remains incomplete. Therefore, the MHD models are still subject to limitations when predicting the occurrence and evolution of solar flares [42].

The SOC model, proposed by Bak et al. in 1987 [43], explains how dynamic systems can spontaneously reach a critical state without external regulation, exhibiting $1/f$ noise and fractal structures, demonstrating scale invariance. The essential idea is that small disturbances can trigger large-scale chain reactions. SOC models are characterized by following a power law distribution, which is a

characteristic that shows up in the distribution of flare power. The SOC model has been applied to solar flare studies. Lu and Hamilton [44] proposed that flares are "avalanches" of small magnetic reconnection events, with flare size linked to the number of reconnections. Karakatsanis and Pavlos [45] further explored the coexistence of SOC and chaotic dynamics in solar activity. Their analysis of the sunspot index revealed significant evidence for the existence of a physical process in solar activity, which corresponds to the existence of a self-organized critical state according to the SOC theory. In addition, Li et al. [46] investigated the relationship between magnetic field parameters and both the intensity and eruptiveness of solar flares, providing important insights for understanding the physical mechanisms of flares and enhancing forecasting capability. A specific example of predicting solar flares based on the underlying physical laws would be a study done by Kusano et al. [47] which presents a physics-based model that predicts large solar flares through a critical condition of magnetohydrodynamic instability, triggered by magnetic reconnection. While these models provide valuable insight, they fall short on the required time frame and accuracy needed for a constant 24-hour prediction.

1.3.2 Statistical models

Statistical models look for strong correlations between solar flares and different parameters. Generally, these models have lower performance but higher interpretability. One of the first statistical relations found in solar flares was the correlation to solar spots. Giovanelli et al. [48] in 1939 demonstrated that flare probability was positively correlated with sunspot group area, type, developmental stage, and growth rate. McIntosh [49] made use of this correlation and created the still widely used McIntosh classification, where classes show a high correlation to flare probability. This relation between sunspot classification and flare occurrence and intensity has been used for prediction by different groups. Bloomfield et al. [50] evaluated flare rates for each McIntosh class, estimated Poisson probabilities for various flare magnitudes, and evaluated different validation approaches. Kangjin Lee et al. [51] use sunspot classification and change in sunspot properties to predict solar flares. Through this approach, the group was able to statistically demonstrate that magnetic flux and its emergence enhance the occurrence of major solar flares. Finally, a different example from Georgoulis et al. [52], which uses a singular calculated parameter B_{eff} that is defined as the effective connected magnetic field strength, shows a high correlation to flare probability.

Statistical methods prioritize interpretability, which often comes at the cost of lower performance. Additionally, it limits the scalability in complex data and generally requires manual feature extraction, which limits the ability for automated real-time prediction.

1.3.3 Machine learning models

Classic machine learning, where feature extraction is still done manually, has been used extensively in solar flare prediction. Generally, the results have been better than traditional statistical methods. Some examples are listed below.

Song et al. [53] used the ordinal logistic regression method to establish a prediction model, which estimates the probability for each solar active region to produce X-, M-, or C-class flares during the next 1-day time period. This is done based on 3 parameters derived from the Solar and Heliospheric Observatory (SOHO), Michelson Doppler Imager(MDI) magnetograms.

Nishizuka et al. [54] compared three different algorithms, Support Vector Machines, k-Nearest Neighbors, and Extremely Randomized Trees, to predict the maximum class of flares occurring in the following 24 hr. Key features, including flare history, magnetic neutral line length, unsigned magnetic flux, and UV brightening, were found crucial for accurate predictions. While this study found k-Nearest Neighbors to perform the best, another study by Bringewald et al. [55] found k-Nearest Neighbors to perform worse compared to the Random Forest and XGBoost models. This underlines one of the problems in solar flare prediction. Different studies use different metrics and different datasets, which creates difficulty when comparing models. The study by Bringewald et al. illustrates the problem with the high dimensionality of solar flare data. Generally, k-Nearest Neighbors (KNN) suffers heavily from high dimensionality and thus will struggle with complex data.

A study by Goodwin et al. [56] evaluated machine learning-based solar flare prediction models in simulated environments using the Space Weather Analytics for Solar Flares dataset. The study found that given enough data, the training window time could be static and still maintain performance; however, it was also found that the solar activity cycle has a crucial influence on model performance.

Overall, machine learning methods have made notable progress in solar flare prediction, but challenges remain. Most importantly, feature selection is an important part when using traditional machine learning methods. This can be quite cumbersome, as it requires time and expertise to manually find patterns in complex data. One way to overcome this problem is the use of deep learning methods, which perform feature extraction as part of the process. This reduces the need for manual feature extraction and improves the generalization when applied to large data sets.

1.3.4 Deep Learning

Deep learning is a subsection of machine learning that has been widely applied in solar flare prediction. Compared to traditional machine learning methods, deep learning methods exhibit a stronger ability to automatically extract features. This trait allows these models to more accurately capture the underlying patterns of a large amount of raw data [57]. Deep learning models come in the following general categories: neural network (NN)-based, convolutional neural network (CNN)-based, recurrent neural network (RNN)-based, hybrid, and transformer-based models.

One prominent example is the development of the Deep Flare Net by Nishizuka et al. [58, 59]. Models like this make use of the automatic feature extraction that comes with deep learning methods. The input data consists of observation images taken by the Solar Dynamic Observatory (SDO). Because of this automatic feature extraction, which is inherent to deep learning approaches, the model can work on live data predicting the probability of flares for the next 24-hour window. While a pure neural network like the Deep Flare Net already improves the prediction capability, more complex models like CNNs can help identify spatial patterns. CNNs have been used to directly analyze solar magnetograms, and because of their increased capability to identify spatial patterns such as magnetic field configurations and polarity inversion lines, which are closely related to flare occurrence. Huang et al. [60] demonstrated that a CNN-based architecture can identify regions of interest in magnetogram data and have comparable performance to high-performing models. While CNNs improve the spatial feature extraction, they struggle with temporal features. Solar flares are strongly linked to the buildup of magnetic energy over time. RNNs are structured in a way more suited

for temporal feature capture. Long short-term memory-based structures are a common variation of RNNs. Jiao et al. [61] proposed a hybrid LSTM-based deep learning framework that integrates regression and classification tasks. For the first time, this approach directly employs time series of SDO/HMI SHARP parameters to continuously predict the peak intensity of solar flares within the next 24–48 hours. To make use of the benefits of both architectures, hybrid structures are developed. One such example is Tang et al. [62], who developed the combination of a CNN and bidirectional Long Short Term Memory (LSTM) to predict whether sunspot groups would produce C and M flare events within the next 24 or 48 hours. They demonstrated that a hybrid model, which integrates the strengths of the individual architectures, outperforms traditional statistical forecasting methods as well as any single machine learning approach.

Transformers provide another approach to capturing temporal dependencies in solar flare prediction. Unlike RNNs, which process sequences recurrently, Transformers use self-attention to identify relationships between observations at different points in time, allowing them to capture both short- and long-term dependencies [63]. Kaneda et al. [64] developed the Flare Transformer, which combines solar magnetograms and sunspot physical features to predict solar flares using attention-based temporal modeling. Similarly, Abdullah et al. [65] proposed SolarFlareNet, a Transformer-based framework that uses time series of SDO/HMI SHARP parameters to predict flares over 24–72-hour windows. These studies demonstrate that Transformers can effectively model the temporal evolution of solar active regions and provide competitive performance for flare forecasting. More recent work has also explored Transformers for detailed event prediction, including the timing and intensity of individual flares [66].

1.3.5 Overview

The previous sections discussed the main categories of solar flare forecasting models, ranging from physics-based approaches to modern deep learning methods. Each category has distinct strengths and weaknesses, making them suitable for different forecasting scenarios. While recent research has largely focused on improving predictive performance, practical forecasting systems must also consider factors such as the effort required to develop and maintain the model and the ability to understand the reasoning behind its predictions. As a result, there is no single forecasting approach that is optimal in every aspect.

To compare the different forecasting approaches, three qualitative criteria are considered: accuracy, operational efficiency, and interpretability. Accuracy refers to the general predictive performance reported in the literature for short-term solar flare forecasting. Operational efficiency describes the ease with which a forecasting system can be developed, deployed, and maintained. It considers computational requirements, the need for manual feature engineering, model updates, retraining, and the amount of domain expertise required to operate the model effectively. A higher operational efficiency indicates that the forecasting system requires less effort to maintain and operate. Finally, interpretability indicates how easily the model’s predictions can be explained in terms of physical processes or observable solar parameters. A highly interpretable model allows researchers to understand why a prediction was made, whereas models with low interpretability often function as black boxes.

The ratings presented in Table 4 are qualitative and represent the general characteristics of each

model category rather than individual implementations. Since performance depends on the dataset, forecasting horizon, and evaluation metric, the values should be interpreted as an overall comparison of the approaches discussed throughout this chapter.

Table 4: Qualitative comparison of solar flare forecasting model categories.

Model Type	Accuracy	Operational Efficiency	Interpretability
Physics-based (MHD, SOC)	Medium	Low	High
Statistical	Low	Medium	High
Traditional Machine Learning	Medium	Low	Medium
Deep Learning	High	Medium	Low

Physics-based models provide the greatest physical insight because they are directly derived from the governing equations describing solar plasma and magnetic fields. Their predictions can therefore be related to physical mechanisms, resulting in high interpretability. However, these models require computationally intensive simulations and expert knowledge to construct and maintain, leading to low operational efficiency. Although they successfully reproduce many aspects of solar magnetic field evolution, their forecasting accuracy remains limited by the incomplete understanding of the mechanisms responsible for flare initiation.

Statistical models are comparatively simple and computationally inexpensive. Their predictions are based on explicit statistical relationships between observable solar parameters and flare occurrence, making them highly interpretable. Nevertheless, they rely on manually selected features and are generally unable to capture the complex nonlinear behaviour of active regions, resulting in lower predictive accuracy. While maintaining these models is less demanding than physics-based approaches, periodic recalibration and feature selection still require human intervention, giving them a medium operational efficiency.

Traditional machine learning models improve upon statistical methods by learning nonlinear relationships from historical data. They generally achieve higher predictive accuracy but still depend heavily on manually engineered features. Selecting, extracting, and validating these features requires significant domain expertise and ongoing maintenance, which reduces their operational efficiency. Their interpretability is lower than that of statistical models but remains moderate, as many machine learning algorithms provide measures such as feature importance that offer insight into the prediction process.

Deep learning models currently provide the highest predictive performance for solar flare forecasting by automatically learning complex spatial and temporal representations directly from raw observational data. The ability to perform automatic feature extraction substantially reduces the manual effort required during model development and maintenance compared to traditional machine learning approaches. Although training deep learning models requires considerable computational resources and large datasets, automated training pipelines can simplify deployment and model updates, resulting in a medium operational efficiency overall. Their primary disadvantage is their limited interpretability, as the learned representations are difficult to relate directly to physical processes.

1.4 Interpretability in Deep Learning

1.4.1 Why interpretability matters.

Interpretability of deep learning models is not just a problem in solar flares but a wider question in applications of deep learning. It is particularly problematic when the model is applied to a scientific problem, since the objective is not only to predict an event but also to establish whether the prediction is based on physically meaningful information. A model may, for example, exploit correlations that are specific to a dataset rather than relationships that are relevant to the underlying physical system. Consequently, predictive performance alone does not demonstrate that a model has learned a physically meaningful representation of the problem [67].

For scientific applications, interpretability can therefore be understood as the ability to relate the information used by a model to quantities or relationships that have a physical meaning. This is stronger than simply explaining an individual prediction. In particular, post-hoc explanations of a black-box model do not necessarily reveal the actual mechanism by which the model makes its prediction [67]. Ideally, physical knowledge should therefore be incorporated into the model itself, while post-hoc analysis can be used to investigate and validate the resulting behaviour.

1.4.2 Physics informed deep learning models

Traditional physics simulations struggle with 3 key problems: high computational cost, especially when considering many-query problems; feasibility, certainly considering high-dimensional problems; and inability to obtain solutions when the understanding of the physics is incomplete. While deep learning models, applied to systems with much available data, don't suffer from these specific issues, they lack the amount of interpretability physical simulations offer. Additionally, there is no guarantee that these models follow the underlying physics at all. An approach to try and combine the strength of these different model types was introduced by Raissi et al. [68]. By introducing a loss function based on the physics involved in the problem, the neural network is trained on finding the pattern within the data and simultaneously on following the known physical laws. This is known as a physics-informed neural network, or PINN. Later studies by, for example, Moseley et al. [69] demonstrated that even training the network only on physical constraints and limited boundary information allowed it to accurately predict the solution far outside the domain where training data existed.

In the original PINN framework, the physical term represents governing differential equations. This approach is particularly powerful when the governing equations of the system are known and can be evaluated during training [68]. In many observational problems, however, the complete physical equations and state variables are not available. In such cases, a strict PINN formulation is not necessarily possible. Nevertheless, the underlying principle remains useful: known physical relationships can be incorporated into the loss function as soft constraints, thereby restricting the solutions that the neural network can learn.

1.4.3 Physical interpretability in solar-flare prediction

The situation in solar-flare prediction differs from the problems for which conventional PINNs were originally developed. The full physical process leading to a flare involves complex magnetohydrodynamic evolution and magnetic reconnection, while the available observations provide only a partial description of the solar atmosphere. There is therefore no single known differential equation that can simply be inserted into the loss function and used to constrain an LSTM forecasting model in the same manner as a conventional PINN.

Nevertheless, this does not mean that physical information cannot be incorporated into the learning process. The distinction is that the available information is generally empirical and phenomenological rather than a complete governing equation. Known relationships between magnetic properties and flare occurrence can therefore be used as softer physical constraints.

In this work, this idea is applied to an LSTM-based flare-prediction model. Rather than optimizing the model exclusively according to the prediction error, an additional loss term is introduced to encourage consistency between the model’s predicted flare probability and the cumulative evolution of SHARP magnetic parameters. The underlying assumption is that temporal changes in photospheric magnetic properties associated with flare productivity contain information about the evolving likelihood of flare occurrence.

This is not a PINN in the strict sense of enforcing a known governing physical differential equation. Instead, it can be viewed as a physics-informed or physically constrained recurrent neural network, in which established solar-physical relationships are incorporated into the optimization objective. The approach follows the broader philosophy of PINNs—restricting the hypothesis space using physical knowledge—while adapting it to a domain in which the complete governing equations are not available in an appropriate observational form.

This distinction is important. The physical constraint does not claim that the cumulative SHARP parameters constitute a physical equation governing flare initiation. Rather, it encodes the weaker but still useful assumption that the temporal evolution of magnetic properties known to be associated with flare productivity should contain information about the evolution of the predicted flare probability. The neural network is therefore encouraged to learn a relationship that is simultaneously predictive and consistent with established empirical knowledge.

The physical constraint can subsequently be complemented by feature analysis. For example, feature importance or attribution methods can be used to determine which SHARP parameters contribute most strongly to the prediction and whether the learned relationships are consistent with previous flare-forecasting studies. Agreement between the physically constrained model and independent feature-analysis methods provides an additional means of assessing whether the model is learning physically meaningful information.

1.5 Chapter overview

These are the central points of this chapter:

1. Solar flares are large eruptions of electromagnetic radiation from the Sun that can have signif-

icant adverse effects on Earth and space-based technologies. Since electromagnetic radiation from a flare reaches Earth in approximately 8 minutes, accurate and timely forecasting is essential to provide sufficient warning and enable effective mitigation measures.

2. Accuracy, operational efficiency, and interpretability are three important qualities of solar-flare forecasting models. While traditional and statistical approaches can offer greater interpretability, deep-learning models can capture complex nonlinear relationships and achieve high predictive performance, but often at the cost of reduced interpretability.
3. Physics-informed neural networks provide one of the most established approaches for incorporating physical knowledge into deep-learning models. However, their conventional formulation requires explicit governing equations, which are often unavailable for complex observational problems such as solar-flare prediction. In such cases, physical interpretability can instead be pursued by incorporating empirically established physical relationships into the model architecture or loss function, complemented by post-hoc feature and attribution analysis.

2 Research Proposal

2.1 Knowledge gap

Despite substantial progress in solar-flare forecasting, an important gap remains in developing models that simultaneously achieve high predictive performance and incorporate physically meaningful constraints when explicit governing equations are unavailable. Deep-learning approaches are well suited to capturing the complex, nonlinear relationships present in solar observations, but their data-driven nature makes it difficult to establish whether the learned representations are consistent with known physical relationships. Physics-informed neural networks offer a framework for addressing this issue, but their conventional formulation relies on explicit governing equations, limiting their direct applicability to observational problems such as solar-flare prediction, where a complete set of governing equations is not readily available.

This creates a need for alternative approaches to physics-informed learning that can incorporate established empirical relationships without requiring a fully specified physical model. To the best of our knowledge, limited work has investigated whether an empirically motivated temporal relationship between SHARP magnetic parameters and flare probability can be incorporated directly into the training objective of a recurrent solar-flare forecasting model. This work addresses this gap by introducing a physics-informed loss function that constrains the relationship between the cumulative evolution of physically relevant SHARP parameters and the flare probability represented by the model’s learned hidden state. Rather than enforcing explicit physical equations, the proposed approach uses an empirically motivated relationship as a constraint on the learned representation, providing a way to introduce physical knowledge into a deep-learning forecasting model when a complete governing system is unavailable.

2.2 Research objectives, questions, and hypotheses

The overarching research objective is to investigate whether physically motivated constraints can be incorporated into a deep-learning solar-flare forecasting model through the loss function, thereby improving the physical interpretability and generalization of the model without substantially reducing its predictive performance.

In particular, this research investigates whether an empirically motivated relationship between the cumulative evolution of physically relevant SHARP parameters and flare occurrence can be used to constrain the learned representation of a probabilistic forecasting model. Unlike conventional physics-informed neural networks, the proposed approach does not require a complete set of explicit governing equations. Instead, physical knowledge is incorporated through relationships that are supported by theoretical considerations and empirical evidence.

To address this research objective, the following research questions are formulated:

1. Which SHARP parameters and derived physical quantities provide the most relevant physical information for solar-flare forecasting?
2. How can established physical relationships between the selected parameters and solar-flare occurrence be incorporated into the loss function of a probabilistic deep-learning forecasting model?

- (a) Which formulation of the physical loss function best represents the identified physical relationship while remaining compatible with the model’s probabilistic output?
 - (b) To what extent does constraining the learned representation using the physical loss function result in physically coherent model behavior?
3. What effect does the physics-informed loss function have on the predictive performance of the forecasting model compared with an otherwise identical data-driven model?
- (a) How does the physics-informed model compare with the standard model in terms of appropriate solar-flare forecasting performance metrics?
 - (b) Does the predicted flare probability of the physics-informed model exhibit a stronger and more physically consistent relationship with the selected physical parameters than the standard model?

The following hypotheses are made based on existing information:

1. SHARP parameters and derived physical quantities that quantify the magnetic complexity, free magnetic energy, magnetic shear, and electric-current structure of active regions contain significant information about subsequent solar-flare occurrence, and models based on these physically motivated quantities will provide greater predictive information than models based on less physically relevant SHARP parameters.
2. Incorporating an empirically supported relationship between the cumulative evolution of physically relevant SHARP parameters and flare occurrence as an additional term in the training loss will constrain the learned representation of the probabilistic forecasting model toward physically consistent behavior without preventing the model from learning predictive relationships from the observational data.
 - (a) A physical loss function formulated to constrain the predicted flare probability according to the empirically observed relationship between cumulative physical parameter evolution and flare occurrence will provide a more effective and physically meaningful constraint than formulations that impose the relationship directly on intermediate model representations or use an incompatible deterministic target.
 - (b) The physics-informed model will exhibit a stronger and more consistent relationship between its predicted flare probability and the selected physically relevant SHARP parameters than the otherwise identical data-driven model, indicating that the physical loss function induces more physically coherent learned behavior.
3. Incorporating the physics-informed loss function will improve or maintain the predictive performance of the forecasting model relative to an otherwise identical data-driven model, while simultaneously increasing the physical consistency of its predictions.
 - (a) The physics-informed model will achieve predictive performance that is equal to or better than that of the standard data-driven model across appropriate probabilistic solar-flare forecasting metrics, with no statistically significant degradation in predictive skill.
 - (b) The predicted flare probabilities of the physics-informed model will exhibit a stronger and more physically consistent relationship with the selected SHARP parameters and their cumulative evolution than those of the standard data-driven model.

2.3 Planning

The work is divided into 7 work packages and their respective sub-packages:

1. **Literature review:** The literature review consists of reading and evaluating many different sources to determine the state of the art and identify the gaps and possible solutions.
2. **Dataset construction:** The dataset construction will start with the identification of relevant data and preparation for training. Based on the available data, a suitable overarching model architecture is chosen, and a labeling strategy is made and executed.
 - (a) **Exploration:** Explore what datasets are available and why one should be chosen over the other.
 - (b) **Input data selection and construction:** Construct the input data for the models and describe each parameter inside the set
 - (c) **Label data selection and construction:** Construct a database of solar flare events and link them to the input data appropriately.
3. **Physical loss function construction:** The dataset construction will give good insight into the available data. Based on this research will be conducted into how a loss function can be constructed which makes use of this data and enforces coherence to the physical phenomena underlying solar flare occurrence. This work package is extended to overlap with the model construction to allow the function to fit properly with the model structure and design.
 - (a) **Research:** Research will be conducted into what physical processes are important for solar flares and can potentially be used in a loss function.
 - (b) **Testing:** Different loss functions will be created and tested while constructing the model to try and achieve the best variation.
4. **Model construction:** The model construction is the longest work package as it will involve creating, training and reiterating on the model based on outcome. It has a slight overlap with model evaluation to allow for tweaks to the model based on results.
5. **Model evaluation:** The model will be evaluated with the chosen metrics and compared to the chosen base model and other similar models in literature.
6. **Finalization:** This work package is mainly meant for writing and defense preparation, but does allow for some overrun of other tasks.
7. **Writing:** The writing will be done throughout the process. Each section in the final report will be written during its corresponding work package.

2.4 Requirements

The requirements below are derived directly from the research questions and are intended to make the research objective experimentally testable.

2.4.1 Overarching requirements

1. At least two forecasting models shall be constructed:
 - a standard data-driven model; and
 - a physics-informed model based on the standard model with the addition of a physical loss term.
2. The standard and physics-informed models shall use the same architecture, input data, pre-processing, labels, training procedure, and hyperparameter settings wherever possible. The physical loss term shall be the only intentional model difference.
3. Both models shall be capable of running on a RTX4060 Ti, training fully within 30 minutes.
4. Both models shall produce a probabilistic prediction of whether an active region will produce at least one M-class or higher flare within the subsequent 24-hour forecasting window.

2.4.2 Dataset requirements

5. The dataset shall cover at least 11 consecutive years and therefore at least one complete solar cycle.
6. The input data shall consist of SHARP summary parameters available at or close to real-time conditions, sampled at a cadence of no more than 60 minutes.
7. Each input sample shall be associated with a clearly defined 24-hour forecasting window.
8. Flare labels shall be binary:
 - positive if the associated active region produces at least one M-class or higher flare within the subsequent 24 hours;
 - negative otherwise.
9. Flare events shall be directly linked to the corresponding SHARP observations through the NOAA active-region number.
10. The dataset construction shall prevent information from the future forecasting window from entering the input features.
11. The division between training, validation, and test data shall be defined before final model evaluation to prevent data leakage.

2.4.3 Physical-parameter requirements

12. The parameters used in the physical loss function shall be selected based on:
 - physical relevance to solar-flare production;
 - evidence from the scientific literature;
 - statistical relationships observed in the constructed dataset; and
 - suitability for calculating a meaningful temporal or cumulative quantity.

13. The physical parameters used in the loss function shall be available from the same input data used by the forecasting model and shall not require additional physical information during operational inference.
14. The physical relationship implemented by the loss function shall be explicitly defined and justified using theoretical reasoning and empirical evidence.
15. The physical loss shall operate on quantities available from the model’s learned representation and/or probabilistic output without requiring the model to directly predict an unobserved physical state.
16. The contribution of the physical loss shall be controlled by an explicitly defined weighting parameter so that its influence on the total training objective can be evaluated.
17. At least one suitable baseline formulation of the physical loss and, where feasible, alternative formulations shall be evaluated to determine whether the observed effect is dependent on the particular loss formulation.

2.4.4 Standard-model requirements

18. The standard model should achieve a True Skill Statistic (TSS) of at least 0.7 on the held-out test set, provided that this target is consistent with the selected model architecture and comparable literature.
19. The standard model shall be sufficiently competitive with existing solar-flare forecasting approaches to provide a meaningful baseline for evaluating the effect of the physical loss.
20. The standard model shall be evaluated using metrics appropriate for imbalanced probabilistic forecasting, including at least TSS, precision and recall.

2.4.5 Physics-informed-model requirements

21. The physics-informed model shall be architecturally identical to the standard model.
22. The physics-informed model shall use the same input data and labels as the standard model.
23. The only intentional modification between the standard and physics-informed models shall be the addition of the physical loss term.
24. The effect of the physical loss shall be evaluated separately from the effect of changes in model architecture or input features.
25. The physics-informed model shall be evaluated using the same test set and evaluation procedure as the standard model.

2.4.6 Computational requirements

33. Both models shall be trainable within 30 minutes on a personal computer containing a RTX4060 Ti.

34. The computational cost of training the standard and physics-informed models shall be recorded and compared.
35. The inference procedure of both models shall not require additional physical simulations or computationally expensive physical calculations beyond the information already available in the forecasting input.

2.5 Chapter overview

1. The overarching research objective is to investigate whether physically motivated constraints can be incorporated into a deep-learning solar-flare forecasting model through the loss function, thereby improving the physical interpretability and generalization of the model without substantially reducing its predictive performance.
2. 7 work packages have been defined consisting of: Literature review, Dataset construction, Physical loss function construction, Model construction, Model evaluation, Finalization, and Writting.
3. The overarching most important requirement is that 2 working models will be constructed, one baseline and one with a custom physics informed loss function. Both models shall be ran under identical conditions on identical data to ensure a fair comparison.

3 Dataset Construction

Two sources of data are combined. The primary magnetic field data comes from the Near Real Time (NRT) SHARP series. Each entry corresponds to a single active region (AR) observation at one-hour cadence and contains parameters computed from the cylindrical equal-area (CEA) projection of the Stokes-inversion vector magnetogram. Flare event data are taken from the Heliophysics Event Knowledgebase (HEK) and cover the same 2013–2025 interval.

3.1 Magnetic field parameters

The Helioseismic and Magnetic Imager (HMI) data products include a set of derived quantities known as Space-weather HMI Active Region Patches (SHARP) parameters. These parameters summarize the magnetic properties of an active region. Rather than using every pixel of the vector magnetogram directly, SHARP parameters compact the information into a few scalar quantities that characterize magnetic flux, gradients, electric currents, helicity, magnetic free energy, and field geometry. Because these properties are closely related to the build-up and release of magnetic energy in solar active regions, they have become standard features in solar flare forecasting studies [36].

The simplest quantity is the total unsigned magnetic flux (USFLUX), which measures the total flux emerging from the active region. Large values of magnetic flux generally correspond to larger and more magnetically complex active regions capable of storing substantial magnetic energy.

Several SHARP parameters quantify the spatial variation of the magnetic field. MEANGBT, MEANGBZ, and MEANGBH represent the mean gradients of the total, vertical, and horizontal magnetic field components, respectively. Strong magnetic gradients are typically found along polarity inversion lines where opposite magnetic polarities meet. These regions are often associated with magnetic reconnection and are therefore considered important indicators of flare emergence.

Electric currents are estimated from the vector magnetic field through Ampère’s law. The mean vertical current density (MEANJZD) describes the average current density perpendicular to the solar surface, while TOTUSJZ measures the total unsigned vertical current integrated over the active region. Large electric currents indicate that the magnetic field deviates from a potential, current-free configuration and therefore contains stored magnetic energy.

Magnetic twist and helicity are represented by MEANALP, MEANJZH, TOTUSJH, and ABSNJZH. The parameter MEANALP estimates the average force-free twist parameter, providing a measure of how strongly magnetic field lines wind around one another. MEANJZH and TOTUSJH quantify the current helicity density and total unsigned current helicity, respectively, while ABSNJZH measures the absolute net current helicity.

The parameter SAVNCP measures the net electric current within each magnetic polarity. Instead of allowing positive and negative currents to cancel, it sums the absolute values of the net current in each polarity separately, making it another indicator of the electrical complexity of the active region.

Magnetic free energy is approximated through MEANPOT and TOTPOT, which estimate the excess magnetic energy relative to a corresponding potential magnetic field. A potential field

represents the minimum-energy state for a given magnetic flux distribution. Consequently, these parameters provide an estimate of the magnetic energy available to power eruptive events such as solar flares and coronal mass ejections.

The geometric relationship between the observed magnetic field and the corresponding potential field is described by MEANSHR and SHRGT45. MEANSHR is the average shear angle between the observed and potential magnetic fields, while SHRGT45 gives the fraction of the active-region area where the shear angle exceeds 45° . Large shear angles indicate that the magnetic field has been significantly stressed, a characteristic frequently associated with flare-producing active regions.

Finally, several parameters describe the size of the active region itself. AREA_ACR gives the deprojected physical area of the active region after correcting for projection effects near the solar limb, whereas SIZE_ACR represents the projected area of the active region in the image plane. NPIX simply records the number of pixels belonging to the HARP mask and serves as a measure of the region’s spatial extent in the original magnetogram.

Together, these SHARP parameters provide a compact yet comprehensive description of the magnetic state of a solar active region. They capture both the strength of the magnetic field and its departure from a simple potential configuration, making them valuable predictors for statistical flare forecasting and machine learning applications. They are all displayed in Table 3.

The data for each AR is sorted by time, and a few filters are applied. First, each region must have a non-zero NOAA AR number. This number is necessary to link flares to the region, so any region without it cannot be accurately labeled. Secondly, a limb filter is applied. This means any AR located more than 70° from the center is filtered out of the dataset. This is done based on the findings by Bobra et al. [70]. It was found that the distortion effect was too large, causing models to perform worse when trained on these samples. Finally, some parameters have gaps of missing values. To deal with these gaps linear interpolation is used, according to the following formula:

$$\hat{x}_{r,t}^{(j)} = x_{r,t_0}^{(j)} + \frac{t - t_0}{t_1 - t_0} \left(x_{r,t_1}^{(j)} - x_{r,t_0}^{(j)} \right) \quad (1)$$

Where x is each single sharp parameter, and t_1, t_0 are the nearest non-missing values. Where if $t_1 - t_0 > 5$ instead of interpolation the entire region is filtered out to avoid creating long sequences of artificial data. Finally, the parameters are created into 24-hour sequences, with the first full sequence of an active region being defined by a full sequence filled with only 0’s for each parameter. This is done so predictions can be made from the first hour of the region’s existence and not just starting at the 24th hour.

In addition to the original features, delta features are also computed. This is done in the following way:

$$\Delta x_{r,t} = x_{r,t} - x_{r,t-1}, \quad \Delta x_{r,t_0} = 0 \quad (2)$$

For the delta features, the padding zeros are explicitly ignored, and the first delta feature after the padding sequence is set to zero to avoid artificial spikes.

3.2 History paramters

For the solar flare data, the same region number filter is applied first. This means the NOAA AR number must be non-zero to properly associate the flare with an active region. Additionally, each flare must also have a flux value, and thus a class associated with it, to assure proper use in the dataset.

In addition to the sharp parameters, 2 additional types of parameters are created. First, based on the work by Liu et al. [71], 15 parameters based on flaring history are created. For each class in (B, C, M, X), an individual decay feature is constructed using the following formulas:

$$\text{Bdec}(x_t) = \sum_{f_j \in F_B} e^{-\frac{t-t(f_j)}{\tau}} \quad (3)$$

$$\text{Cdec}(x_t) = \sum_{f_j \in F_C} e^{-\frac{t-t(f_j)}{\tau}} \quad (4)$$

$$\text{Mdec}(x_t) = \sum_{f_j \in F_M} e^{-\frac{t-t(f_j)}{\tau}} \quad (5)$$

$$\text{Xdec}(x_t) = \sum_{f_j \in F_X} e^{-\frac{t-t(f_j)}{\tau}} \quad (6)$$

Where F_B (F_C , F_M , F_X , respectively) represents the set of B-class (C-class, M-class, X-class, respectively) flares that occurred in the same AR before the data observation time point t , $t(f_i)$ denotes the occurrence time of the flare f_i , and τ is a decay constant that is set to 12 following the findings from Jonas et al. [72]. t is the current time, and t_k is the time the flare occurred. In addition, a total energy decay value is used, both directly and in its logarithm. Given in the following equations:

$$\text{Edec}(x_t) = \sum_{f_j \in F} E(f_j) \cdot e^{-\frac{t-t(f_j)}{\tau}} \quad (7)$$

$$\log\text{Edec}(x_t) = \sum_{f_j \in F} \log(E(f_j)) \cdot e^{-\frac{t-t(f_j)}{\tau}} \quad (8)$$

Where $F = F_B \cup F_C \cup F_M \cup F_X$ and $E(f_i)$ is the magnitude of flare f_i . To complete the history features, 9 direct features are used to express how many flares occurred before the current time. Bhis, Chis, Mhis, and Xhis are the total number of flares of that class that occurred in the same AR before the time point t_i . Additionally, each of these parameters has a corresponding 1-day feature. Bhis1d, Chis1d, Mhis1d, and Xhis1d are the number of flares of the class occurring within the last 24 hours. And finally, Xmax1d gives the maximum flux value that occurred in the last 24 hours.

3.3 Solar phase constant

The sun follows a roughly 11-year-long cycle, which strongly affects the amount of active regions and solar flares. To give the model information on how far along the sample is in the solar cycle, 2 parameters are constructed. First, a phase is computed in the following way:

$$\phi(t) = \frac{(t - t_{\text{ref}}) \bmod P_{\odot}}{P_{\odot}} \quad (9)$$

Where t_{ref} is the last element of the following set of known solar minima: (1996-09-01, 2008-12-01, 2019-12-01) and $P_{\odot} = 4015 \approx (11 \text{ years})$. This phase is encoded using the following formulas:

$$[\sin(2\pi\phi(t)), \cos(2\pi\phi(t))] \in [-1, 1]^2 \quad (10)$$

This results in two parameters *solarphase_sin* and *solarphase_cos*.

3.4 Normalization

For normalization, standard Z-score normalization statistics are calculated from the training set. The padding zeros are again specifically ignored during the normalization process. This is especially important for parameters such as USFLUX, which generally has a magnitude in the order of 10^{21} or more, which means a large amount of artificial 0's will strongly affect the resulting normalized parameter. The mean and standard deviation are calculated using the following formulas:

$$\mu_j = \frac{1}{|T_{\text{train}}|} \sum_{(i,t) \in V_{\text{train}}} X_{i,t}^{(j)} \quad (11)$$

$$\sigma_j = \sqrt{\frac{1}{|V_{\text{train}}|} \sum_{(i,t) \in V_{\text{train}}} (X_{i,t,j} - \mu_j)^2 + \epsilon} \quad (12)$$

Where V_{train} is the total training set excluding all the padding zeros. For the standard deviation, $\epsilon = 10^{-8}$ is used to ensure no division by 0 occurs. Then the normalized features are calculated using:

$$\tilde{X}_{i,t,j} = \frac{X_{i,t,j} - \mu_j}{\sigma_j} \quad (13)$$

The normalization statistics are saved and are used on any testing or validation set to avoid data leakage.

3.5 Label construction

A sliding-window approach is used to construct the input sequences and prediction targets for the LSTM model. For each time step, the model receives the preceding 24 hours of observations as an input sequence. This 24-hour window is then used to predict whether at least one solar flare of class M or X will occur during the following 24 hours. The window is subsequently shifted forward in time, and the process is repeated to generate overlapping input sequences and their corresponding labels. Each sequence is assigned a label of 0 if no M- or X-class flare occurs in the subsequent 24-hour prediction window, and a label of 1 if at least one such flare occurs.

3.6 Dataset Split

The final dataset is split into 3 parts. A training set of 60%, a validation set of 20%, and a testing set of 20%. This split is done based on AR numbers to avoid overlapping sequences in different sets. This does have the advantage over a time-based split of a roughly equal division of samples from each period in each set. Because solar activity is cyclical, the number of active regions and the proportion of flares greatly differ each year. By random division of active regions over the full available timeframe, this effect is largely negated.

3.7 Chapter overview

1. Solar flare and magnetic field data between 2013-2025 is downloaded from the HMI and HEK database.
2. Filtering is applied for missing entries and regions more than 70° from the center.
3. For each SHARP feature a corresponding delta feature is created.
4. Using the flare data, history features are constructed based on the work of Liu et al. [71]. And a feature representing the solar phase is constructed using sinusoidal encoding.
5. Each feature is normalized using Z-score normalization.
6. Sequences are constructed with padding 0's and labels using a sliding window approach, where a M or X flare in the next 24 hour means label 1 and 0 otherwise.
7. The dataset is split in a training, validation and testing set using a 60%, 20% and 20% split based on active regions.

4 Methodology

The research objective is to improve the physical interpretability of deep learning models while maintaining high performance. The designed model follows a Long short term memory or (LSTM) architecture. In addition to the standard LSTM model, a custom loss function is created to try and make the model more physically consistent and interpretable.

4.1 Model design

Studies show that LSTM models perform well in prediction of solar flares based on the sharp parameters [73]. An LSTM model is a version of a recurrent neural network (RNN). It tackles the vanishing gradient problem that makes standard RNNs struggle with learning long-range dependencies [74]. The key idea is the introduction of a memory cell to carry information over long sequences. It does this using three gates to manage the flow of information: The forget gate, the input gate and the output gate. The general model architecture is depicted in Figure 5.

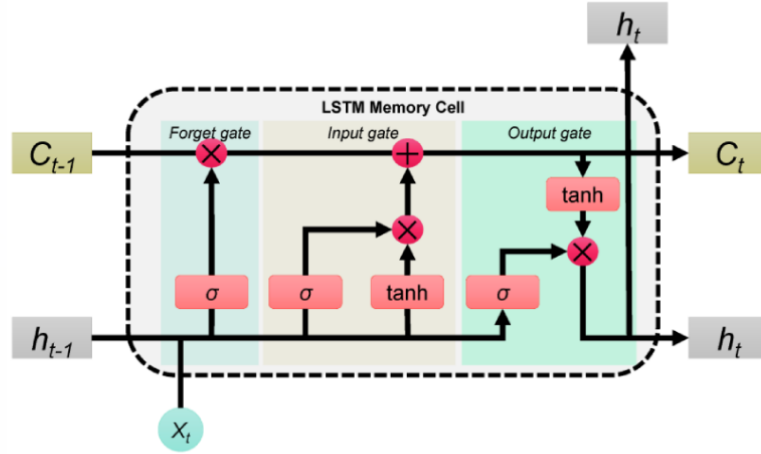


Figure 5: LSTM model architecture overview [75]

The forget gate (Equation 14) determines which information from the previous cell state should be retained or discarded.

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (14)$$

where f_t is the forget gate activation, h_{t-1} is the previous hidden state, x_t is the current input, W_f and b_f are the learnable weight matrix and bias, and $\sigma(\cdot)$ denotes the sigmoid activation function.

The input gate (Equation 15) determines how much new information should be written to the cell state.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (15)$$

where i_t is the input gate activation, and W_i and b_i are the corresponding learnable parameters.

The candidate cell state (Equation 16) contains the new information that may be added to the cell state.

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (16)$$

where $\tilde{c}_t$ is the candidate cell state, W_c and b_c are the learnable weight matrix and bias, and $\tanh$ is the hyperbolic tangent activation function, producing values between -1 and 1 .

The cell state is updated by combining the retained information from the previous cell state with the newly generated candidate information (Equation 17).

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (17)$$

where c_t is the updated cell state, c_{t-1} is the previous cell state, and $\odot$ denotes element-wise multiplication.

The output gate (Equation 18) determines which information from the cell state is exposed as the hidden state.

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (18)$$

where o_t is the output gate activation, and W_o and b_o are the corresponding learnable parameters.

The hidden state (Equation 19) is computed by filtering the updated cell state through the output gate.

$$h_t = o_t \odot \tanh(c_t) \quad (19)$$

where h_t is the hidden state at time step t , which is passed to the next LSTM cell and can also serve as the network output.

For the specific application, the model receives a tensor of shape $B \times T \times P$ with batch size $B = 64$, sequence length $T = 24$ and parameter count $P = 55$ consisting of 19 sharp features, 19 delta features, 15 history features, and 2 solar phase encodings. The LSTM consists of two stacked layers with a hidden size of 32 units. The first LSTM layer processes the input sequence and produces a sequence of hidden states, which, after Dropout ($p=0.5$) during training, are passed as input to the second LSTM layer. This second layer generates a refined sequence of hidden states that captures increasingly abstract temporal patterns. For classification, only the final hidden state, h_T , is used by the classification head, as it represents a learned summary of the complete input sequence. Before classification, the final hidden state is passed through layer normalization, which normalizes the feature activations across the hidden dimensions and applies learnable scaling and shifting parameters to improve training stability. Dropout is subsequently applied to reduce overfitting by randomly masking 50% of the features during training. The resulting feature vector is then processed by a two-layer feedforward network. The first layer reduces the dimensions from

32 to 16 using a ReLU activation function and a dropout of half the standard, so 0.25, while scaling the survivors by 4/3. The second is a simple linear layer that reduces the dimension from 16 to 1 and makes the output a single logit. The entire process follows this equation:

$$\hat{y} = \mathbf{W}_2 \text{ReLU}(\mathbf{W}_1 \text{Dropout}(\text{LN}(\mathbf{h}_T)) + \mathbf{b}_1) + b_2 \quad (20)$$

The final logit $\hat{y}$ gets transformed into a value between 0 and 1, which can theoretically be interpreted as a probability, using the sigmoid function:

$$p = \sigma(\hat{y}) = \frac{1}{1 + e^{-\hat{y}}} \quad (21)$$

This head is subject to the weighted binary cross-entropy loss function:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{B} \sum_{i=1}^B [w^+ y_i \log(\sigma(\hat{y}_i)) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))] \quad (22)$$

Where $\sigma(\hat{y}_i)$ is the final prediction output of the model, which should be as close to the corresponding label y_i as possible. w^+ is the positive class weight, which is set to 28, roughly the fraction of non-flare samples/flare samples.

To attempt to make the model more physically consistent, a separate output is used. Rather than using only the final hidden state, the physics based head operates on the complete sequence of hidden states generated by the second LSTM layer $H = \{h_1, h_2, \dots, h_T\}$. A single linear layer is applied independently to each hidden state to produce a sequence of logits according to the following formula:

$$s_t = W_s \cdot h_t + b_s \quad (23)$$

which are transformed by a sigmoid activation into a sequence of confidence values:

$$\hat{p}_t = \sigma(s_t) \quad (24)$$

These values represent the model's confidence at each timestep rather than only at the end of the sequence. During training, this confidence trajectory is compared with the cumulative sum of a selected SHARP parameter. For a selected SHARP feature $x^{(j)}$, the cumulative signal is computed as:

$$c_t^j = \sum_{T=1}^t x_T^j \quad (25)$$

The agreement between the predicted confidence trajectory and the cumulative physical quantity is measured using the Pearson correlation coefficient:

$$r_P(u, v) = \frac{\sum_t (u_t - \bar{u})(v_t - \bar{v})}{\sqrt{\sum_t (u_t - \bar{u})^2 \cdot \sum_t (v_t - \bar{v})^2 + \epsilon}} \quad (26)$$

It ranges between -1 and 1, where 1 is fully correlated and -1 is inversely correlated. Two other types of correlation are used as well. Spearman correlation, which converts the data values into ranks and then computes the Pearson correlation on those ranks, and Kendall correlation, which

compares the movement of two variables pairwise. These two types of correlation involve discrete rankings and as such are not differentiable. To use them in a loss function, they need to be differentiable, so approximations are used. For Spearman correlation, instead of using hard ranks, a soft approximation using the following formula is used:

$$\tilde{r}_t(u) = 1 + \sum_{t' \neq t} \sigma\left(\frac{u_{t'} - u_t}{\tau_{rank}}\right) \quad (27)$$

Where $\tau_{rank} = 0.05$ functions as a temperature parameter and σ is defined as the sigmoid function. This approximation approaches the real rank as $\tau_{rank} \rightarrow 0$. By calculating the Spearman correlation of this soft ranking, the rough but differentiable Spearman coefficient is found:

$$Spearman(u, v) \approx r_P(\tilde{r}(u), \tilde{r}(v)) \quad (28)$$

For the Kendall correlation, a similar approach is used. Kendall correlation relies on pairwise ranking through the use of a $sign(x)$ function, which is not differentiable at $x = 0$. To combat this, the $sign(x)$ function is replaced with a $\tanh(x/\tau_{rank})$ function that as $\tau_{rank} \rightarrow 0$ approaches $sign(x)$. The formula is given as:

$$\hat{\tau}(u, v) = \frac{2}{T(T-1)} \sum_{t < t'} \tanh\left(\frac{u_t - u_{t'}}{\tau_{rank}}\right) \tanh\left(\frac{v_t - v_{t'}}{\tau_{rank}}\right) \quad (29)$$

Where T is the total number of unique pairs and $\tau_{rank} = 0.05$ again. All three of these correlations fall between -1 and 1, where -1 is completely negatively correlated and 1 is completely positively correlated. By implementing this into a loss function, the correlation gets added as a second goal for the model. The loss function is constructed as the output of the second head of the model, and its correlation to the cumulative sum of a chosen parameter up to that point:

$$\mathcal{L}_{phys}^{(i)} = 1 - Corr\left(\hat{p}_t^{(i)}, c_t^j\right) \quad (30)$$

The gradients created by both these losses flow through the whole model, creating a total loss function that will have the following form:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{BCE} + \alpha\mathcal{L}_{phys} \quad (31)$$

Where α is a term to dictate which loss dominates. It is important to note that both objectives function independently of one another. The main objective is for the model to classify the samples according to the expectation of a solar flare event. This is done through the classification head, which takes in the final hidden state of the LSTM. The physics correlation task tries to align the flow of the hidden states of the full sequence with the flow of a selected parameter. The extent to which both tasks can be achieved depends on to what extent the LSTM can represent a state in which both objectives can be maximized and which is favored through the selection of the value for α .

For the training process, the AdamW optimizer, which decouples the weight decay term from the gradient update, is used:

$$\theta^{(e+1)} = (1 - \lambda)\theta^{(e)} - \eta \frac{\hat{\mathbf{m}}^{(e)}}{\sqrt{\hat{\mathbf{v}}^{(e)} + \epsilon_{Adam}}} \quad (32)$$

Here θ is the model parameters, or the weights, e is the current training step or epoch, $\eta = 8 \cdot 10^{-6}$ is the learning rate, $\lambda = 0.01$ the weight decay coefficient, $\hat{\mathbf{m}}$ and $\hat{\mathbf{v}}$ are the bias-corrected first and second moment estimates, and ϵ_{Adam} a tiny constant added to avoid division by zero. To avoid the gradient exploding, the gradient norms are clipped to not go above 1:

$$\nabla\theta \leftarrow \nabla\theta \cdot \min\left(1, \frac{1.0}{\|\nabla\theta\|_2}\right) \quad (33)$$

As can be seen from the high dropout, high weight decay, and low learning rate. The model was quite prone to overfitting during the training process. An additional method to deal with this problem is the use of early stopping. The most desired metric of True Skill Statistic (TSS) is used as a guide for the most optimal configuration. During the training process, the epoch where the validation set has the highest TSS is chosen as the final epoch.

4.2 Physical loss function motivation

The generally accepted theory behind the cause of solar flares relies on the buildup of energy in the magnetic field, which suddenly releases. Energy gradually builds up in magnetic fields, particularly above active regions where strong, complex magnetic fields emerge through the surface near sunspots. As these magnetic field lines are twisted, sheared, and stretched by motion in the solar plasma, they store energy in a stressed configuration. This leads to the emergence of flux ropes into the corona. Which in turn forces oppositely directed magnetic field lines together beneath it, forming a thin current sheet where magnetic reconnection can occur [76, 77]. Magnetic reconnection is a phenomenon first described by James W. Dungey [78]. Generally plasmas obey the principle known as Alfvén’s theorem also called the frozen-in flux theorem, which states that in the limit of ideal magnetohydrodynamics (infinite electrical conductivity), magnetic field lines are ”frozen” into the fluid and move with it, preserving their topology over time [79]. In regions of high magnetic shear, which are also referred to as current sheets following Ampère’s law (Equation 34), Alfvén’s theorem breaks down and the topology of magnetic field lines can change.

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \quad (34)$$

In these regions of high magnetic shear, field lines break and reconnect with other field lines. These new connections exist in a lower energy state, this causes magnetic energy to be converted into heat and high energy particles [80, 81]. The exact mechanism and process remains to be fully understood.

The key physical theory for this study is that flare occurrence should be preceded by and associated with conditions for buildup of magnetic stress, elevated free energy, and strong current density in the active region. A model that predicts high flare probability during periods of low magnetic stress, or low probability during periods of extreme magnetic stress, is physically inconsistent with this theory. The physical loss function is designed to punish this.

4.3 Evaluation metrics

Solar flare prediction models require very specific evaluation metrics to deal with 2 important characteristics of the problem. First, predicting flares correctly is much more important than predicting no flare correctly. In addition, the dataset suffers from large class imbalance, where over

95% of the samples belong to the negative class. This makes accuracy a very misleading metric, as just predicting the negative class will result in a very high accuracy score. The first metrics that show a much more useful picture are precision and recall.

$$Precision = \frac{TP}{TP + FP} \quad (35)$$

$$Recall = \frac{TP}{TP + FN} \quad (36)$$

Where recall is more important than precision as it shows the fraction of solar flare events the model predicted correctly. However, optimizing for either of these metrics will cause the model to perform in an unbalanced manner where either false positives or false negatives are less relevant. A much more balanced score which is also very commonly used in space weather forecasting is the True Skill Statistic.

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN} \quad (37)$$

This metric takes both false positives and false negatives into account, ensuring a balanced score which does not favor one over the other. In addition, it is not sensitive to class balance. In solar flare prediction this is very important, not only because the dataset is very imbalanced. When comparing model performance on different parts of the solar cycle the class balance will change, it is important that the metric which is used is not sensitive to this. So, for model optimization TSS will be used. State-of-the-art comparable models in the literature report TSS values in the range of 0.6 to 0.8, with recall around 0.9 and precision around 0.2. These values serve as the performance reference for the standard model.

Because the model outputs a continuous probability, a classification threshold must be selected to convert probabilities to binary predictions. An initial threshold of 0.07 is used during training, based on previous trials. At evaluation time, the threshold is optimized over the range 0.01 to 0.99 by selecting the value that maximizes TSS on the validation set. This threshold is then applied to the test set without further adjustment.

4.4 Hyperparameter search

In the section on model construction, a few hyperparameters were given. In this section, the process used to find these hyperparameters is discussed.

For this process the Optuna framework was used [82]. Optuna treats the full training pipeline as a black-box function mapping hyperparameters to validation performance, in this case the TSS metric. The search space includes the following parameters: Hidden size, number of layers, dropout rate, weight decay, learning rate and positive weight. These parameters control the model's capacity, regularization strength, and learning dynamics. Optuna explores the space using a Tree-structured Parzen Estimator (TPE) sampler which, unlike traditional grid or random search, builds a probabilistic model of the hyperparameter space to iteratively predict and evaluate the most promising configurations.

To ensure no data leakage, the test set is not included in the process. Instead, the training and validation sets are combined and split into 5 sets. Again ensuring that no 2 sets contain the same active region number, which causes temporal leakage. By redividing the samples into a few different sets instead of keeping the training and validation set as is, the model parameters are ensured to not hyper-optimize for the original training and validation set that will be used in the training of the final model.

To improve the computational efficiency, a median-based pruning strategy is used. After each cross-validation fold within a trial, intermediate performance is reported. If this falls below the median performance of the previously completed trials, the trial is terminated early. Otherwise, the model is trained identically as it would during training, with early stopping and threshold determination based on the validation part of the new set.

The final objective value for each hyperparameter configuration is the aggregate TSS score result from each set.

4.5 Feature study

Feature selection is commonly used to improve predictive performance, reduce model complexity, and facilitate interpretation of the information used by a predictive model [83]. In this study, feature selection is used to determine which input variables contribute materially to the predictive skill of the baseline model and whether a reduced feature set can preserve or improve forecasting performance. The approach is motivated by previous solar-flare forecasting studies demonstrating that a relatively small subset of physically motivated predictors can provide substantial predictive skill. Bobra and Couvidat [70], for example, applied feature selection to 25 active-region magnetic parameters and found that only a subset was required to obtain good flare-prediction performance. Similarly, Liu et al. [71] developed an LSTM for active-region solar-flare prediction using 40 magnetic and flare-history features and reported that models using only 14–22 of the most important features could outperform models using the complete feature set. These results motivate investigating whether the information contained in the comparatively large feature space used here is redundant for an LSTM, and whether a smaller subset can retain or improve forecast skill.

For the parameter study the dataset follows a similar approach to the hyper parameter study. The training and validation set are used and the testing set is held out. This separation is necessary because repeated model or feature selection on an evaluation set can introduce selection bias and lead to optimistic estimates of predictive performance [84]. Within each elimination step, the 80% tuning pool is further partitioned into seven repeated randomized AR-level train/validation splits. For each split, 75% of the available ARs are assigned to training and 25% to validation. To account for the substantial class imbalance and variation in flare productivity between active regions, ARs are first sorted according to their per-AR flare rate and then randomly assigned to the two subsets. This procedure distributes ARs with different levels of flare activity across the train and validation sets and reduces the sensitivity of feature-importance estimates to the particular composition of an individual split. The seven splits are therefore treated as repeated randomized resampling estimates rather than statistically independent datasets.

Feature selection is performed using sequential backward elimination based on permutation impor-

tance. Sequential backward elimination is a wrapper-based feature-selection approach in which the predictive model is repeatedly evaluated while features are removed according to their contribution to model performance [83]. In contrast to filter methods, which rank variables independently of the predictive model, this approach evaluates feature utility directly related to the performance of the LSTM model.

The procedure begins with all 35 (19 SHARP, 1 solar phase and 15 history) feature groups active. At each elimination step, a separate model is trained for each of the seven AR-level train/validation splits using the current set of active feature groups. For each trained model, a baseline prediction is first obtained on the unmodified validation data and its swept TSS is recorded. The contribution of each active feature group is then estimated using permutation importance.

For a given feature group, all columns belonging to that group are jointly shuffled across the validation samples while the trained model and all other feature groups remain unchanged. The model is then reevaluated on the permuted validation data. Permutation importance measures the change in TSS per feature. The feature group with the lowest mean importance is removed after each iteration. The LSTM is then retrained using the reduced feature set, and the procedure is repeated. Feature groups with large positive importance therefore tend to be retained, whereas groups with importance close to zero contribute little measurable predictive information given the remaining features. Negative importance indicates that, on average, permutation of the group improves validation TSS. Such a result is interpreted as evidence of low or unstable predictive utility and may indicate redundancy, noise, or interactions that are detrimental to the fitted model.

Because feature groups may contain correlated or overlapping information, the resulting elimination order should not be interpreted as an intrinsic physical ranking of the variables. Instead, it represents their conditional marginal predictive utility: the contribution of a feature group given the other groups that remain available to the model at that stage of elimination. This distinction is important because permutation importance can underestimate the importance of a variable when correlated predictors contain overlapping information that remains available after permutation [85]. The group-wise approach consequently provides a more meaningful assessment of the contribution of related physical predictors than independently permuting every underlying column. Elimination continues until only 10 feature groups remain. This 10-group configuration is defined as the minimal surviving feature subset for the study. Once this point is reached, a final evaluation-only pass is performed without further feature elimination. This final pass characterizes the predictive performance of the selected subset without allowing the resulting measurements to influence the feature-selection procedure.

4.6 Chapter overview

1. The model is an LSTM which has a baseline variant and a physics informed variant with a custom loss function
2. Because of the large class imbalance, the chosen evaluation metric for the models is the True Skill Statistic or TSS. Unlike accuracy or the F1-score it is not sensitive to the class imbalance.
3. To study the effect of hyperparameters the Optuna framework is used [82].

4. A separate parameter study is outlined based on the findings by previous solar flare prediction model studies like Bobra et al. [70] and Liu et al. [71], which found that reducing the number of parameters can actually improve the model performance.
5. For both the hyperparameter study and the parameter study, the test set is not included to prevent data leakage.

5 Results

5.1 Hyperparameter tuning

To evaluate the hyperparameters beyond picking the trial with the best TSS score, three complementary sets of plots are created. There is one significant outlier, trial 2, which has a TSS of 0.23. This is not unexpected as early trials have more variance as the optimizer explores aggressively and does not have a good score which it can use as a reference to stop early. However, for figure clarity the total axis is reduced which means it is not included. For the per fold TSS distribution box plot, this axis reduction removes 1 or 2 outliers per fold. A total of 80 trials are started, but because of the early stopping mechanism from Optuna some trials are never finished. This saves computing time, but does reduce the total number of finished trials to 48.

The first plot shows the mean TSS for each completed trial. It shows how the model performance evolves across sequential Optuna trials. Each point corresponds to a full hyperparameter configuration evaluated using the mean cross-validated TSS across the 5 different folds. Displayed in Figure 6 the left plot shows each trial's TSS score.

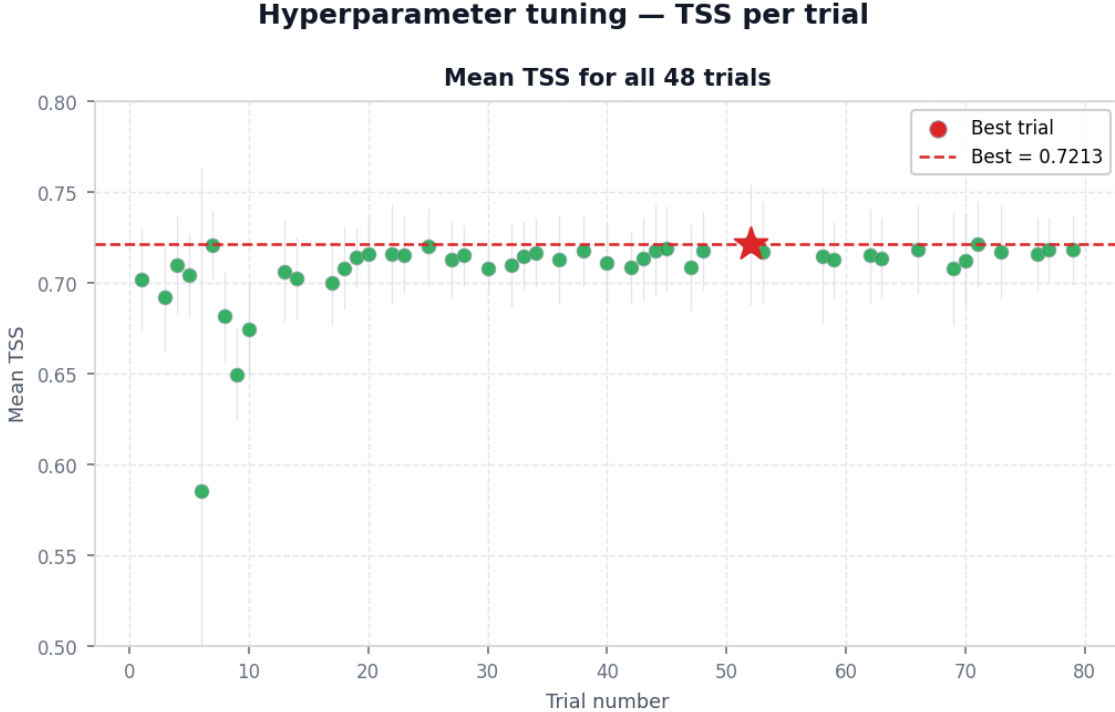


Figure 6: Mean TSS for each completed trial, with TSS on the vertical axis and trial number on the horizontal axis.

The plots show expected behavior from Optuna's TPE sampler, as it progressively concentrates sampling in regions of higher performance [82]. In the earlier trials, the results show large variance

between trials, ranging between a TSS of 0.72 all the way down to 0.58. Around trial 12 the score stops ending lower than 0.70. This is partly because of the early stopping, which can clearly be seen around trial 55 where multiple trials were not completed. The final best trial is 52 but there were others like trial 7 and 71 which come very close. Overall the large variance combined with quick jumps in the TSS score in the first trials, followed by the decrease in variance and few and minor improvements, shows effective navigation of the solution space.

The second set of plots is a set of 6 scatter plots, one for each hyperparameter. In addition, a violin plot is added for the 2 discrete hyperparameters: Number of layers and Hidden size. A Locally Weighted Scatterplot Smoothing (LOWESS) curve is used for one of the discrete hyperparameter: The Dropout rate, and the continuous hyperparameters: Learning rate, Weight decay and Positive weight. This allows for taking a closer look at each individual parameter and their distribution. The first plot is the hidden size depicted in Figure 7.

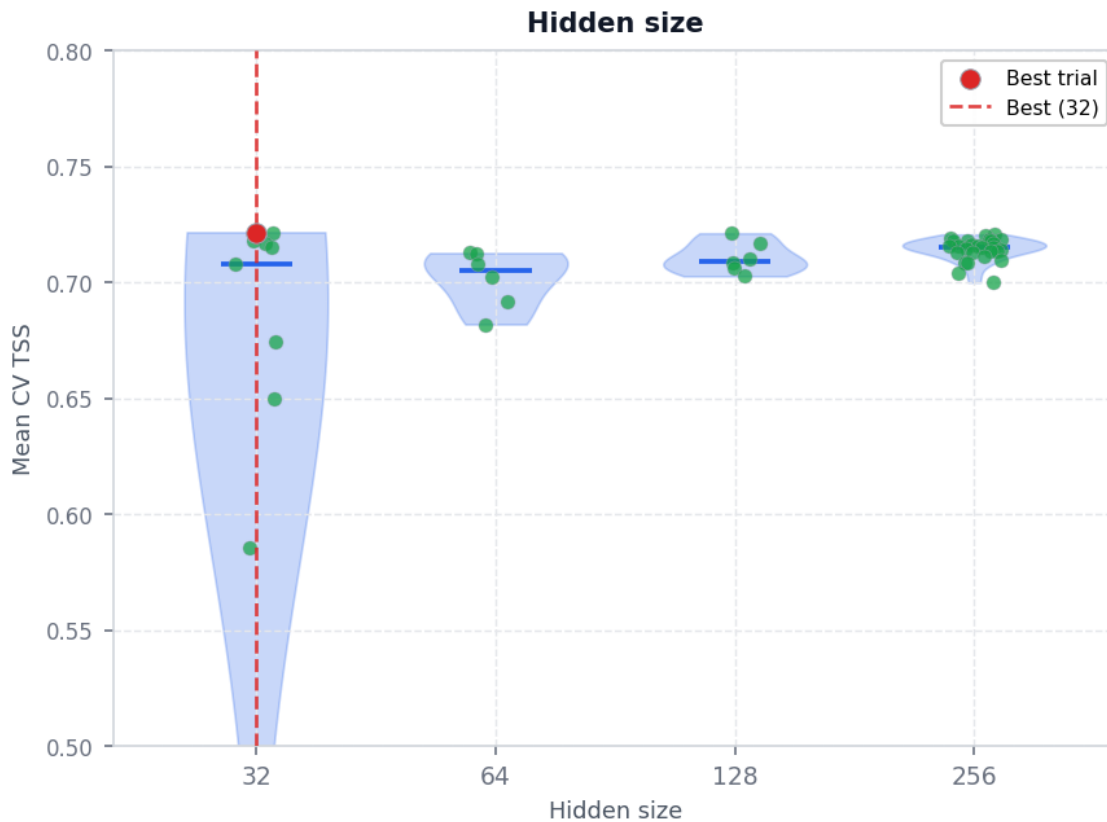


Figure 7: Scatter plot of the each completed trial, with a violin plot distribution on the background. With TSS score on the vertical axis and the associated hidden size on the horizontal axis.

As can be seen, the best trial had a hidden size of 32. However, the highest average TSS score is achieved by the hidden size of 256. This may be cause by the large number of extreme outliers.

The only 3 visible outliers below a score of roughly 0.68 are in the size 32 line. In addition, the strongest outlier of 0.23 is also part of the 32 group. So, while only using the average suggests the best performer is 256, going by the best performer, the highest score was achieved using 32. In the next plot, depicted in Figure 8, a similar pattern appears.

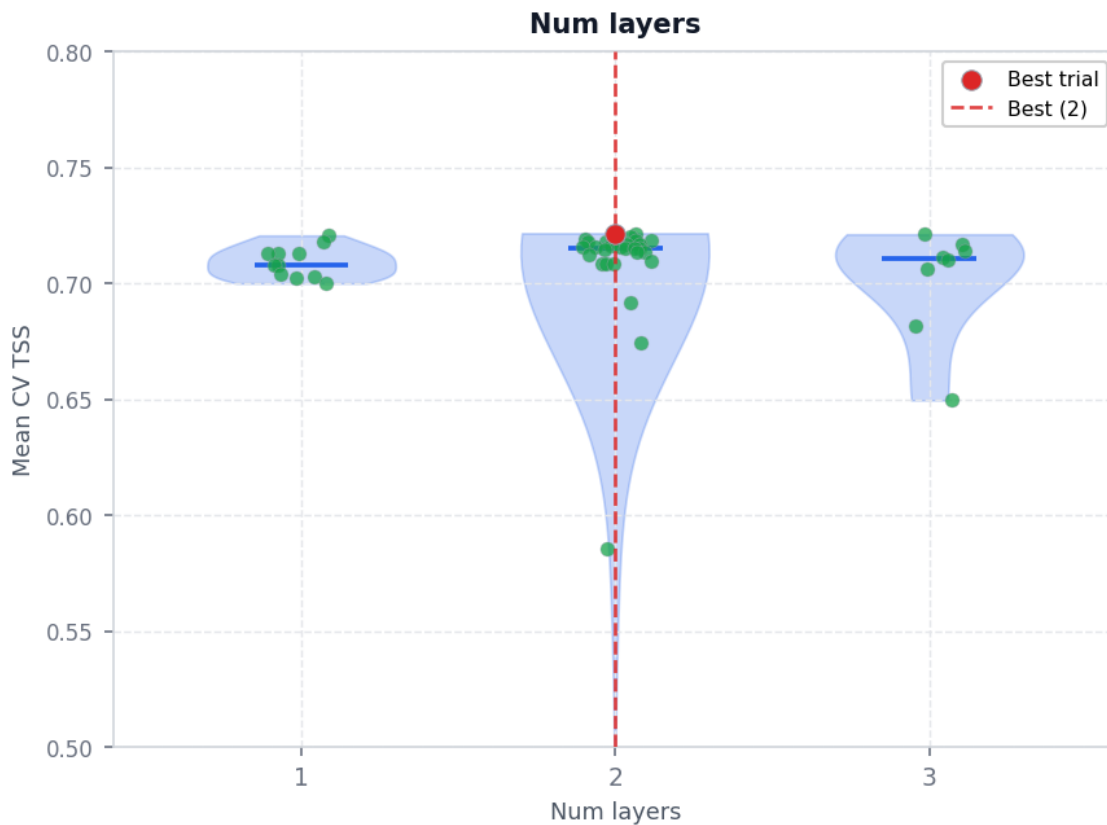


Figure 8: Scatter plot of the each completed trial, with a violin plot distribution on the background. With TSS score on the vertical axis and the associated number of layers on the horizontal axis.

The best trial found has 2 layers for the model size. There are again outliers in this category, but unlike the hidden state, the best average is also found at 2 layers. There is also much more trails in the 2 layer configuration compared to the 1 and especially 3 layers configuration. This shows that the Optuna optimizer has found 2 layers to be optimal and changing this degrades the performance too much. The next hyperparameter is still tested as a discrete parameter, but does use the LOWESS trend line instead of the violin plot as there are too little samples per value. It is displayed in Figure 9.

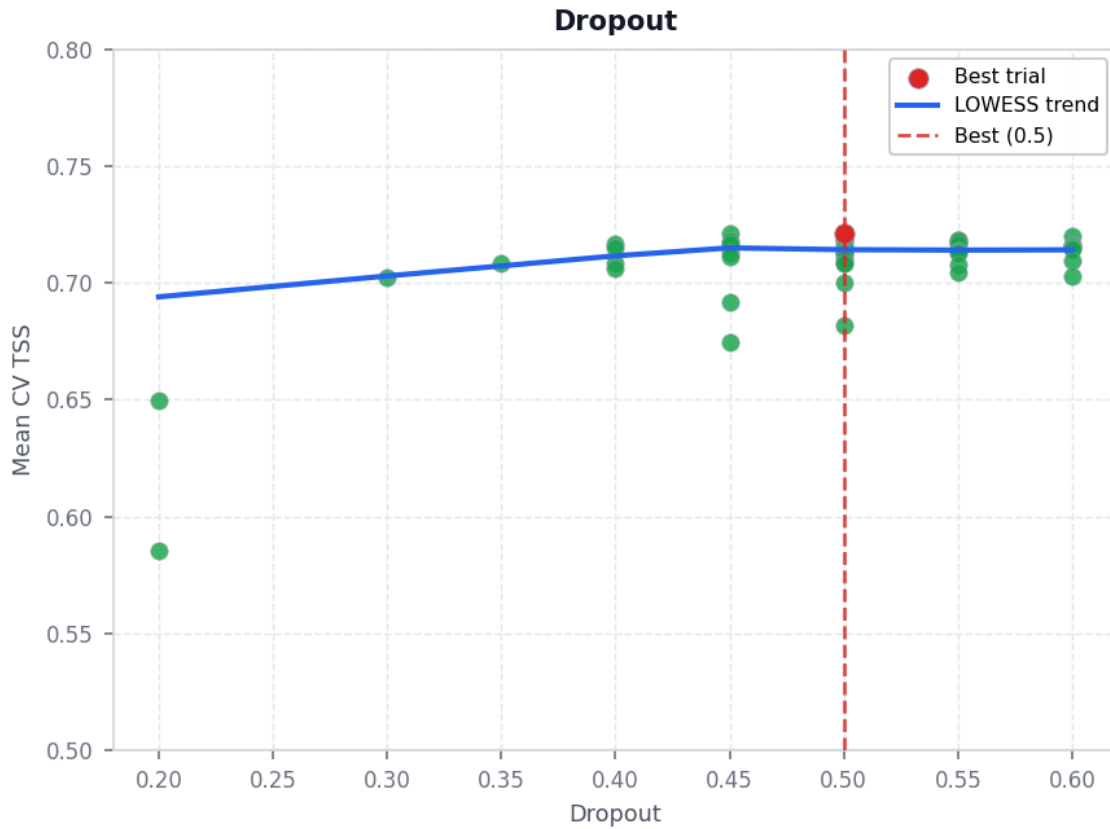


Figure 9: Scatter plot of the each completed trial, with a LOWESS plot distribution on the background. With TSS score on the vertical axis and the dropout rate on the horizontal axis.

Here it is clear that lower values under perform and the optimizer quickly moves to higher values for the dropout rate. Dropout values between 0.45 and 0.60 all have average TSS scores around 0.72, with the best trial having a dropout of 0.50. The next hyperparameters are tested as continuous values. The first being learning rate displayed in Figure 10.

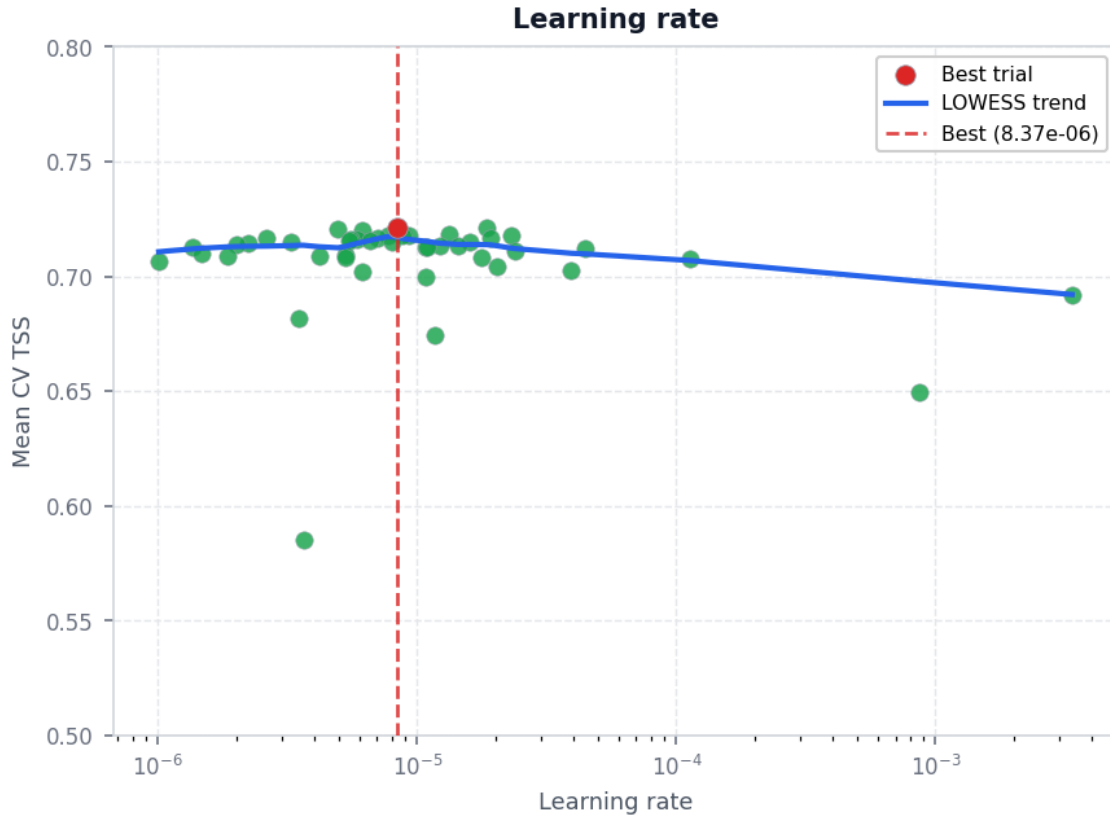


Figure 10: Scatter plot of the each completed trial, with a LOWESS plot distribution on the background. With TSS score on the vertical axis and the learning rate on the horizontal axis, with logarithmic scaling.

With continuous values it is easier to see how the Optuna optimizer operates. It quickly identifies that higher learning rates do not perform well and starts accumulating trials towards the lower end. With only 3 completed trials using a learning rate above 10^{-4} . This means that all the other completed trials fall between 10^{-6} and 10^{-4} , with the highest scoring trial having a learning rate of $8.37 \cdot 10^{-6}$. A similar pattern can be seen in the graph for weight decay, depicted in Figure 11.

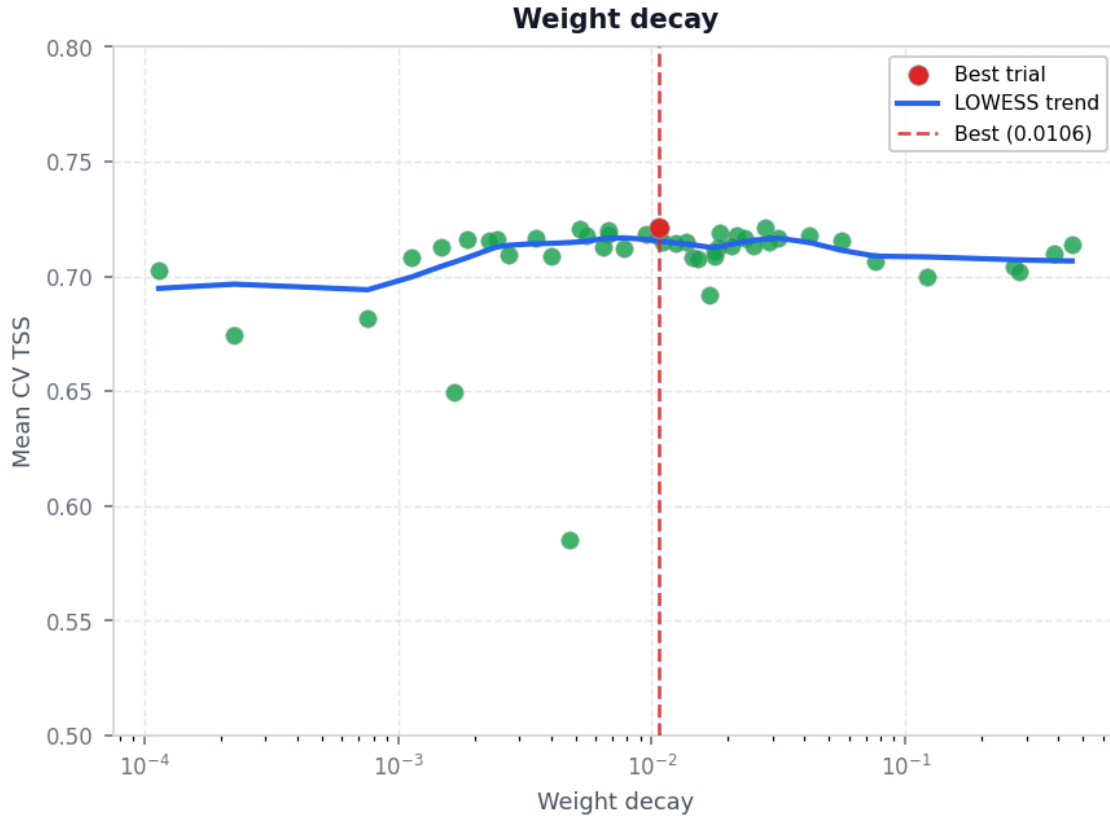


Figure 11: Scatter plot of the each completed trial, with a LOWESS plot distribution on the background. With TSS score on the vertical axis and the weight decay on the horizontal axis, with logarithmic scaling.

While, a similar clustering appears for the weight decay parameter, it is less aggressive compared to the learning rate. Most values lie between 10^{-3} and 10^{-1} with the highest score being achieved in the center at 0.0106. The worst performing completed trials are towards the lower end of the spectrum. This clustering behaviour does not repeat for the last continuous parameter, depicted in Figure 12.

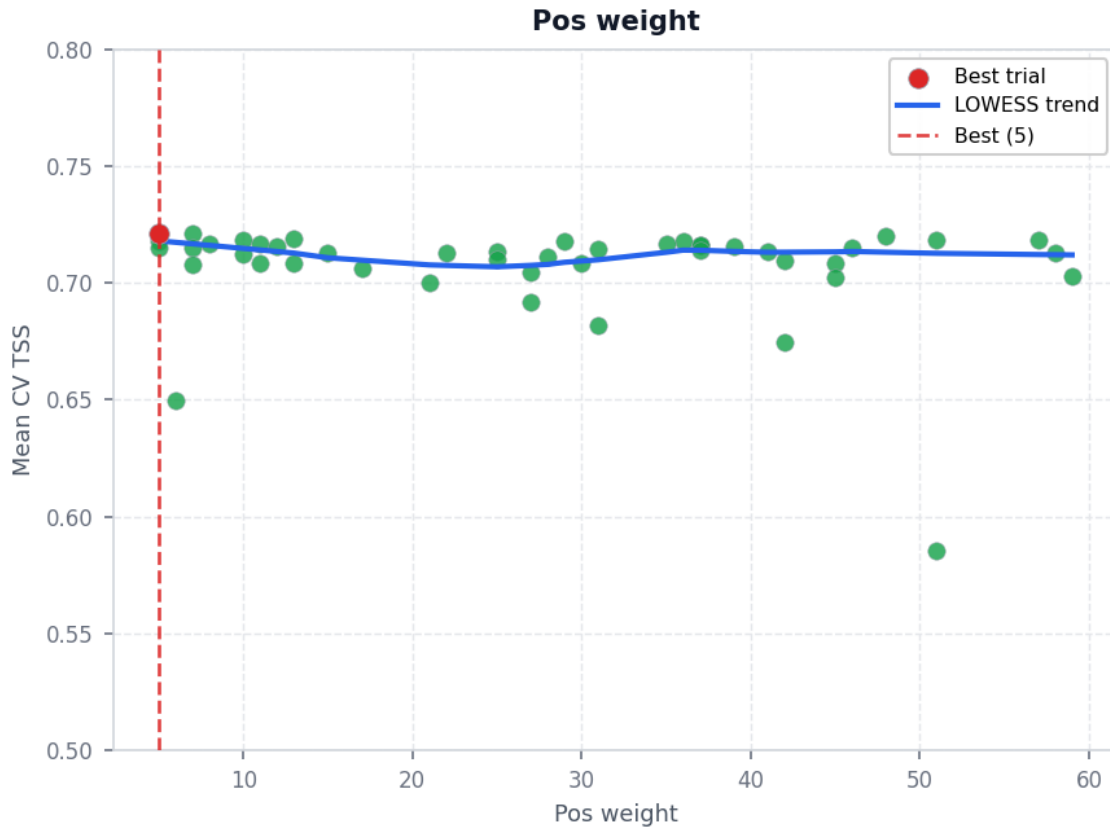


Figure 12: Scatter plot of the each completed trial, with a LOWESS plot distribution on the background. With TSS score on the vertical axis and the positive weight on the horizontal axis.

The positive weight plot shows a roughly even distribution of TSS scores. The best completed trial is found at a positive weight of 5, but other high scores are found at completely different values ranging from 5 all the way to 58. This indicates a lack of sensitivity of the models TSS score to the postive weight parameter.

The final plot provides information on the difference in performance among the 5 splits. The plot shows 5 box plots, one for each split, and the best achieved average. The plot is depicted in Figure 13.

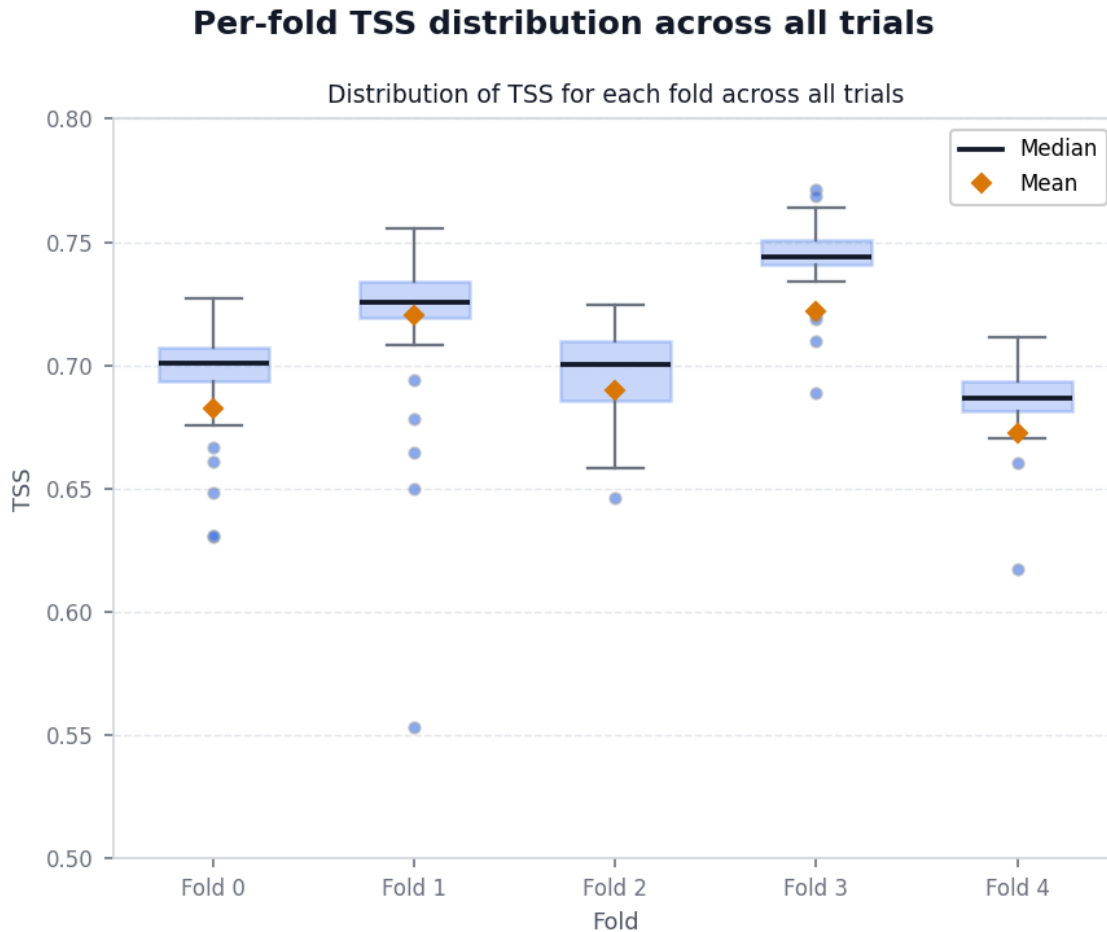


Figure 13: Box plot of TSS score for each of the 5 data splits used. With the fold number on the horizontal axis, the TSS score on the vertical axis, a orange square indicating the mean and a black line indicating the median.

Because of some extreme outliers towards the lower end, the median is significantly higher than the mean for each of the folds. There is a also a clear difference in performance per fold. With fold 3 scoring the best with a median TSS of 0.74 and fold 4 scoring the worst with a median of 0.68.

For each hyperparameter the value from the highest scoring trial is used, except for the positive weight. While the Optuna framework was able to find high performance hyperparameters, TSS score alone is not the full picture. The theoretical impact of the hyperparameters matters too. For the number of layers, dropout, learning rate and weight decay, the highest value was found at the center of a cluster of trials. This indicates that large deviations from the found values could significantly alter performance. However, for the hidden size and positive weight a wide range of values had high scoring trials, indicating less sensitivity of the TSS to these parameters. For

the hidden size, partly because of outliers, the average TSS is higher at larger hidden size. But increasing the hidden size, increases model complexity and thus possibly overfitting. Which based on the high dropout, low learning rate and high weight decay, indicate to be a problem. So, to combat this the hidden size is not altered and kept at 32. The positive weight is different. This parameter has a theoretically correct value of 28 based on the class balance. Because the TSS score does not show high sensitivity to the positive weight parameter, this is changed from 5 to 28, in accordance with the theoretical value. The final resulting hyperparameters are displayed in Table 5.

Table 5: Selected hyperparameters from Optuna tuning and their corresponding values used in final training.

Hyperparameter	Tuned value	Actual value used
Hidden size	32	32
Number of layers	2	2
Dropout	0.50	0.50
Learning rate	8.37×10^{-6}	8.37×10^{-6}
Weight decay	1.06×10^{-2}	1.06×10^{-2}
Pos weight	5.0	28.0

5.2 Baseline model

The baseline model runs with α set to 0. Meaning the model is only affected by the classification loss. The model is trained for a maximum of 100 epochs with an early stopping mechanism that stops after no increase in performance for 30 epochs. The BCE loss throughout the training process is depicted in Figure 14.

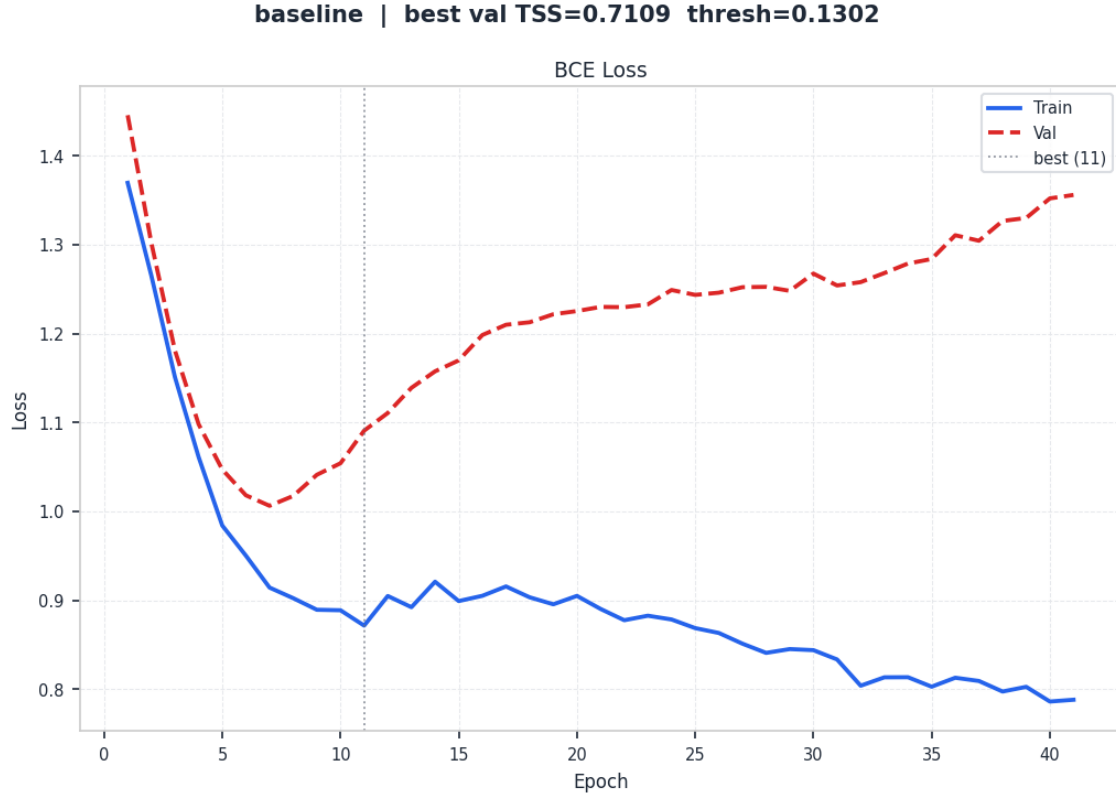


Figure 14: The vertical axis shows the value of the loss and the horizontal axis shows the corresponding epoch. The blue line indicates the training loss and the red line indicates the validation loss. The best epoch, epoch 11, is shown with a dotted gray line.

Based on the BCE loss plot, the model shows clear signs of overfitting. Around epoch 7, the lowest validation loss is hit, while training loss keeps decreasing. While training loss has a slight increase from epoch 10 to 20 it continues decreasing after. Early stopping finds epoch 11 to be the optimal point, as validation TSS was found to be highest here. Notably, the lowest validation loss and highest validation TSS do not occur at the same epoch. Figure ?? shows the evolution of the TSS score for both the fixed 0.07 threshold and the threshold swept on the validation set, which was 0.1302 for the run with the best TSS.

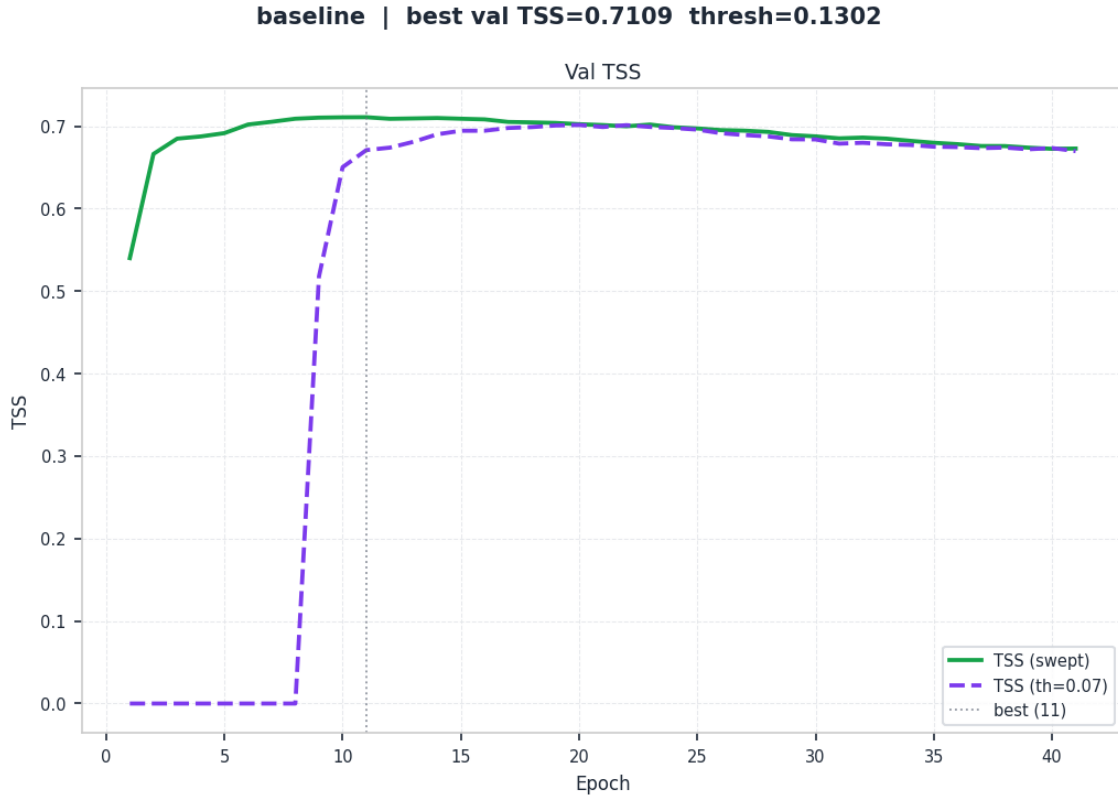


Figure 15: The vertical axis shows the validation TSS and the horizontal axis shows the corresponding epoch. The green line indicates the TSS corresponding to the sweeping threshold and the purple line indicates the TSS for the constant threshold. The best epoch, epoch 11, is shown with a dotted gray line.

Below epoch 8 the model classification is not strong enough to push the sample below the constant threshold, meaning there is no distinction being made, resulting in all samples being labeled as flare and a TSS of 0. The swept TSS score peaks at epoch 11, which means the final model is returned to the state it was in at the end of epoch 11. Interestingly the TSS of the constant threshold and the swept threshold converge around epoch 20. The evolution of the swept threshold is presented in Figure 16.

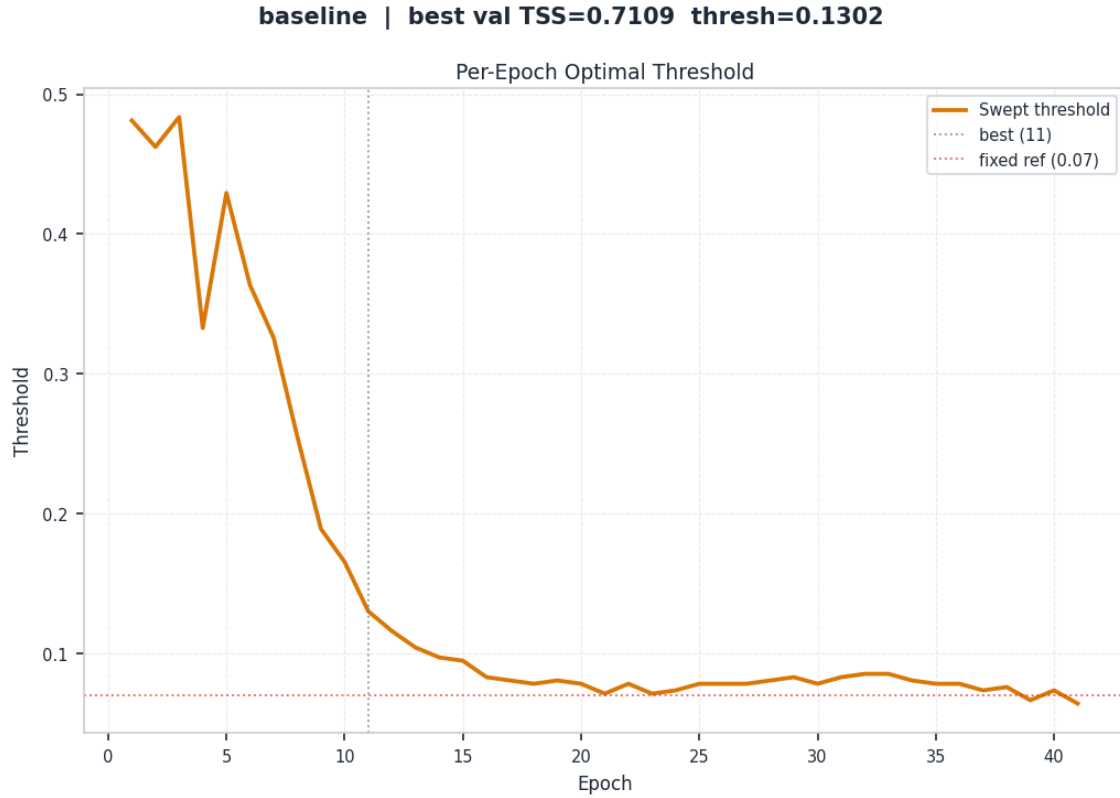


Figure 16: The vertical axis shows the threshold value and the horizontal axis shows the corresponding epoch. The solid orange line indicates the swept threshold value corresponding to the best validation TSS and the dotted red line indicates the constant reference threshold. The best epoch, epoch 11, is shown with a dotted gray line.

It shows that as the model makes a stronger distinction between the 2 classes the optimal threshold decreases, reaching a relatively stable state around epoch 16 of 0.09 close to the reference threshold. As the threshold decreases the amount of false positives increase and the false negatives decrease. Because of the large class imbalance this is a good trade off for the TSS score, as a false negative has much more effect than a false positive. The class separation is clearly shown in Figure 17

Module 2 -- Probability Distributions | baseline

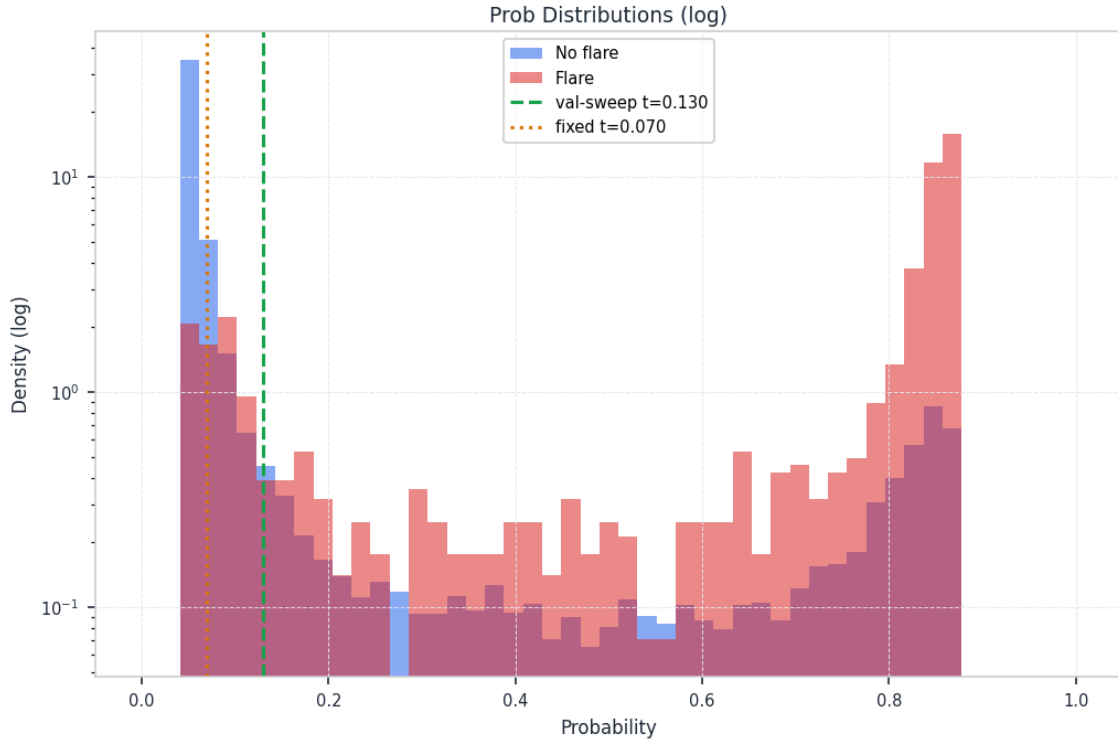


Figure 17: Output of the model found at epoch 11 on the test samples. The vertical axis shows the sample density value on a log scale and the horizontal axis shows the probability score, or the model output. The red indicates a flare sample and the blue indicates a no flare sample, with dark red showing where they overlap. The swept threshold of the best run is shown as a dotted green line and the reference threshold is shown as the orange line.

The figure shows that the negative class is clearly pushed towards and concentrates around 0 and the positive class is pushed towards and concentrates around 1. Notably, non of the samples are fully pushed towards either 0 or 1. This can be explained by the large amount of generalization and few amount of epochs, meaning the model is less inclined to make strong decisions. Another notable point is the mismatch between the swept threshold and the threshold optimal for TSS. Moving the threshold farther left would have the effect of decreasing the amount of flares missed while increasing the amount of false flare prediction significantly. The impact of the threshold selection process can clearly be seen in Figure 18.

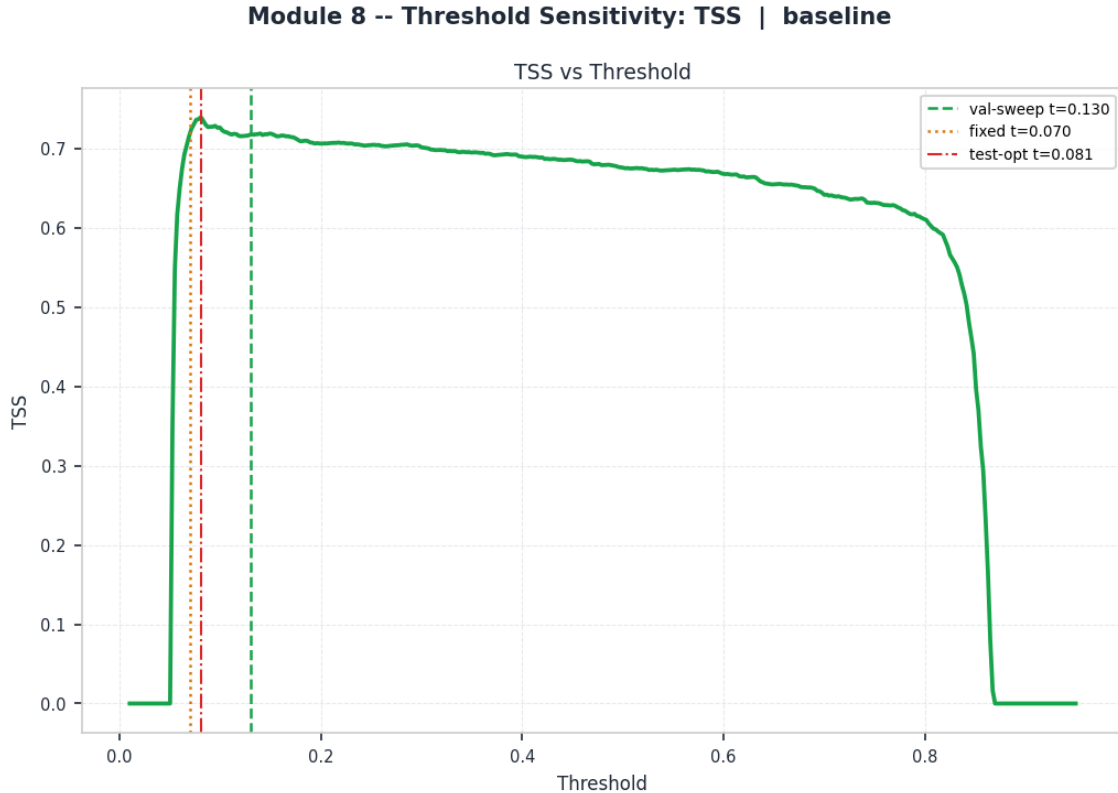


Figure 18: Shows the TSS score of the model from epoch 11 on the test samples for each threshold value. The vertical axis shows the TSS score and the horizontal axis shows the threshold value. The solid green line shows the TSS score for each threshold value, the dotted lines show the validation swept threshold, the reference threshold and the optimal threshold for the highest TSS on the test set, in green, yellow and red respectively.

The first noticeable thing is the misalignment of the both the reference threshold and the swept threshold. While the optimal threshold for the test set would be 0.081 resulting in a TSS of 0.74, the threshold found by the validation sweep of 0.130 results in a score of 0.72. The full confusion matrix corresponding to this threshold is shown in table 6.

Table 6: Confusion Matrix baseline model on test set		
	Predicted Negative	Predicted Positive
Actual Negative	TN = 27252	FP = 4381
Actual Positive	FN = 199	TP = 1185

These values correspond to a TSS score of 0.72, a precision of 0.2129, and a recall of 0.8562. Showing that the model clearly favors not missing flares over predicting flares too often. This is to be expected, based on the strong positive weight that is assigned to the flare class. To highlight

the importance of the evaluation metric the F1 score per threshold is depicted in Figure 19

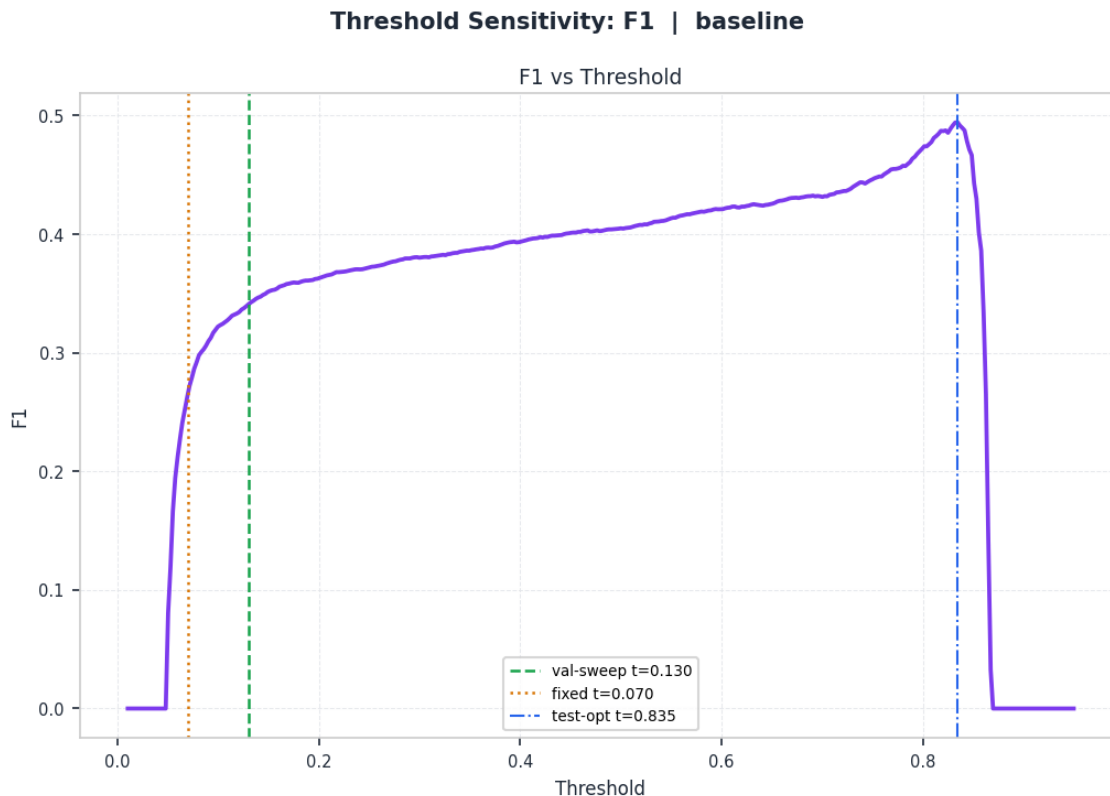


Figure 19: Shows the F1 score of the model from epoch 11 on the test samples for each threshold value. The vertical axis shows the F1 score and the horizontal axis shows the threshold value. The solid purple line shows the F1 score for each threshold value, the dotted lines show the validation swept threshold, the reference threshold and the optimal threshold for the highest F1 score on the test set, in green, yellow and blue respectively.

As can be seen in the figure, if the F1-score were to be chosen as the most optimal performance metric the threshold selection would have varied widely. This is because the F1 score is sensitive to the class imbalance.

While this model does not enforce any correlation between the output of the physics-based head and any parameter through the loss function, the head still exists and generates an output. The strongest Pearson correlation of this output to the cumulative sum of each parameter is displayed in Figure 31, see appendix.

While this model is not subject to any physical loss function, the second head still naturally correlates with some parameters. Most strongly with the TOTPOT or total energy parameter, showing

an average correlation around 0.6. As this model does not enforce any correlation, this is purely coincidental. To inspect the effect of the parameters on the output more, a saliency analysis is conducted using the mean of the gradient times the input. This shows which features the models focus most on. These maps are displayed in Figure 32, see appendix, for the history features, Figure 33, see appendix, for the delta features and solar phase cosine encoding and Figure 34, see appendix, for the SHARP features and solar phase sine encoding.

This map shows expected behavior for an LSTM; the gradient travels backwards from timestep 24 to timestep 1 and gets weaker and weaker throughout for all features. In addition, it shows that the model most strongly relies on direct SHARP parameters instead of delta features or the history features. For the solar phase encoding, the cosine encoding provides a stronger saliency to the model compared to the sin encoding.

Overall, the baseline model shows decent performance with a TSS of 0.72 and a clear preference for recall over precision. The final test is done on the first 5 months of 2026, which is completely unseen and unrepresented data. This is done to check if the model can maintain its performance in operational deployment. The results are given in Table 7.

Table 7: Confusion Matrix baseline model 2026 data using threshold 0.097

	Predicted Negative	Predicted Positive
Actual Negative	TN = 7759	FP = 1684
Actual Positive	FN = 56	TP = 364

The model performance drops somewhat compared to the test set performance. TSS decreases from 0.72 to 0.69 when using the optimal threshold. However, it should be considered that the 2026 set only consists of 9,861 sequences compared to roughly 56,000 in the test set.

5.3 Physics-informed model

For the physics informed model, α starts at 0.9 and decreases by 0.01 each epoch to a minimum of 0. Otherwise the model is trained under exactly the same conditions, with early stopping after 30 epochs without improvement and a maximum of 100 epochs. The TSS difference between the physics constraint model and the base model is given in table 8.

Table 8: Δ TSS (+ better) for different correlation measures.

	pearson	spearman	kendall
USFLUX	-0.0515	-0.0534	-0.0522
MEANGAM	-0.0664	-0.0657	-0.1199
MEANGBT	0.0077	-0.0071	0.0116
MEANGBZ	0.0180	0.0041	0.0140
MEANGBH	-0.0396	-0.0321	0.0177
MEANJZD	-0.0027	-0.0041	0.0091
TOTUSJZ	-0.0578	-0.0401	-0.0531
MEANALP	-0.0170	-0.0104	-0.0582
MEANJZH	-0.0135	-0.0149	-0.0602
TOTUSJH	-0.0555	-0.0558	-0.0535
ABSNJZH	-0.0357	-0.0877	-0.0322
SAVNCPP	-0.0686	-0.0682	-0.0169
MEANPOT	-0.0792	-0.0894	-0.0754
TOTPOT	-0.0157	-0.0151	-0.0032
MEANSHR	-0.0571	-0.0582	-0.1002
SHRGT45	-0.0669	-0.0722	-0.0844
AREA_ACR	-0.0581	-0.0583	-0.0664
SIZE_ACR	-0.0637	-0.0625	-0.0770
RPIX	-0.0707	-0.0651	-0.0213

Baseline TSS = 0.7177

This shows that with the exception of a few almost all models perform slightly worse compared to the baseline. With only 2 pearson, 1 spearman and 4 kendall models performing better than the baseline. MEANGBZ is the only parameter performing better with each correlation type. To investigate further 2 models are presented as case studies: The best model: Pearson MEANGBZ and the worst model: Kendall MEANGAM

5.3.1 Best model case study: Pearson MEANGBZ

The confusion matrix is given in Table 9, showing a lower amount of false negatives compared to the base model, but a larger number of false positives. This shows that will the performance is roughly the same when looking purely at TSS, depending on priorities when model may be preferable.

Table 9: Confusion Matrix Pearson MEANGBZ model with threshold 0.1820

	Predicted Negative	Predicted Positive
Actual Negative	TN = 26611	FP = 5022
Actual Positive	FN = 146	TP = 1238

While the classification loss is much weaker in the beginning due to the high α value it is still

present, the evolution of this loss is displayed in Figure 20

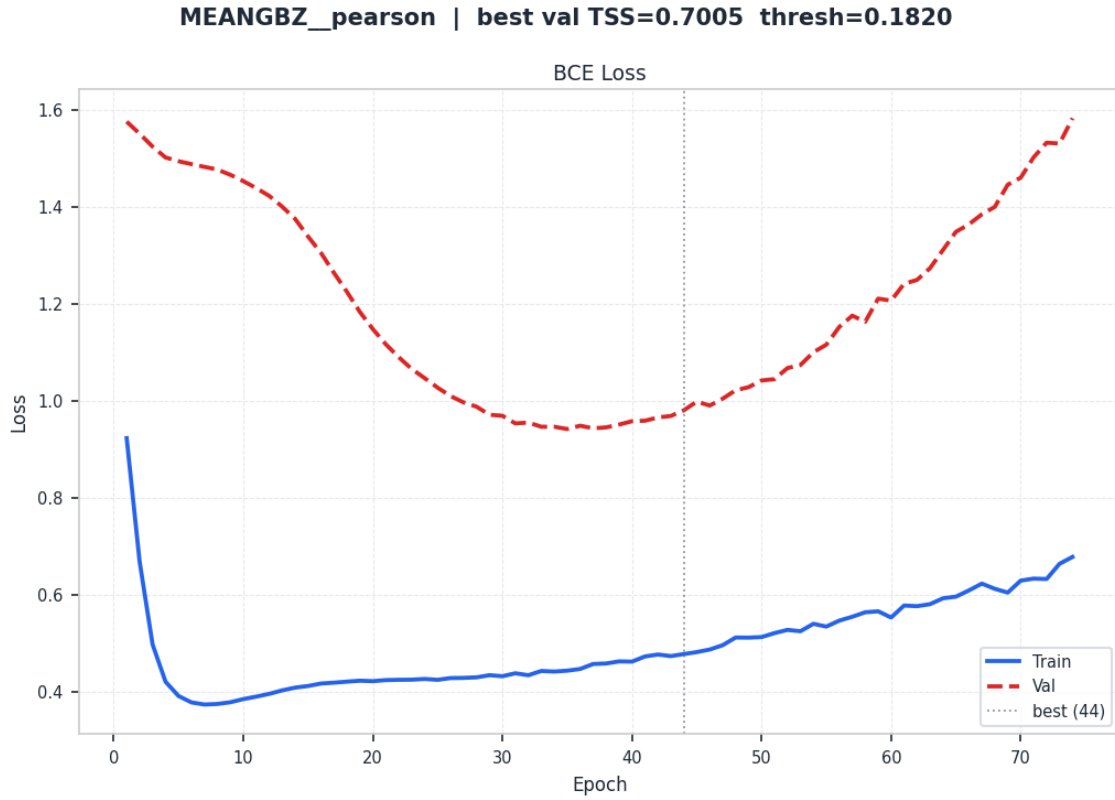


Figure 20: The vertical axis shows the value of the loss and the horizontal axis shows the corresponding epoch. The blue line indicates the training loss and the red line indicates the validation loss. The best epoch, epoch 38, is shown with a dotted gray line.

Compared to the baseline model the training loss is actually lower at 0.4 at epoch 10 compared to the 0.9 of the baseline model. This is because of the scaling factor α , which is 0.8 at epoch 10. Since the loss is scaled with $1 - \alpha$ removing this scaling factor would create a loss of 2 which is much larger than the baseline models loss. Going forward to epoch 50 which is slightly after epoch 44, the epoch used for the final model, α is 0.4 meaning the loss of 0.5 is 1.25 unscaled. This shows that while the loss is growing because of the scaling, the adherence of the model to the training data is improving. As for the physical loss, displayed in Figure 21.

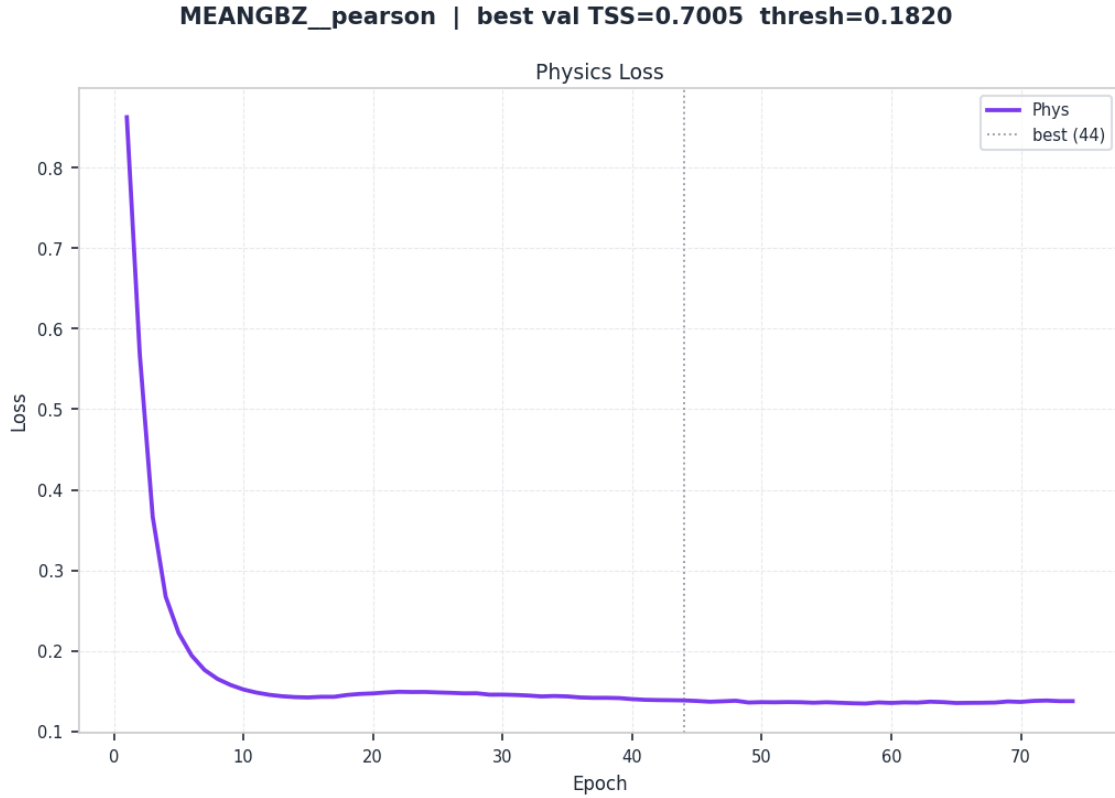


Figure 21: The vertical axis shows the value of the loss and the horizontal axis shows the corresponding epoch. The purple line indicates the physics loss. The best epoch, epoch 38, is shown with a dotted gray line.

When the scaling factor increases, the physics loss stays relatively flat. Comparing epoch 10, where α is 0.8, to epoch 50, where α is 0.4, the scaled physics loss has not changed, indicating that the unscaled physics loss has actually increased. This indicates that as the classification loss gets a larger scaling factor and the physics loss gets a lower scaling factor, the models starts to favor the classification loss and ignoring the physics loss. The TSS performance of the model is depicted in Figure 22.

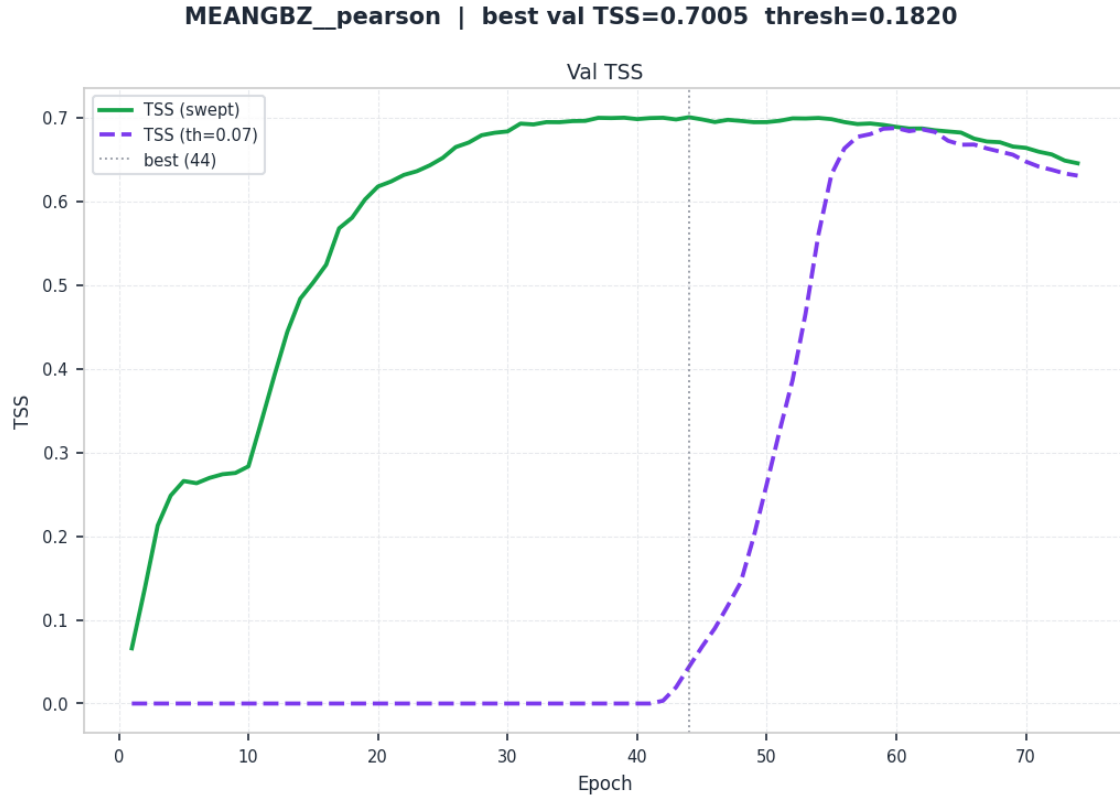


Figure 22: The vertical axis shows the validation TSS and the horizontal axis shows the corresponding epoch. The green line indicates the TSS corresponding to the sweeping threshold and the purple line indicates the TSS for the constant threshold. The best epoch, epoch 38, is shown with a dotted gray line.

The best validation TSS is slightly lower than the validation TSS score of the baseline model, but not by much. The most noticeable thing is the much higher threshold that is found by the validation threshold sweep. Figure 23 shows this effect more clearly.

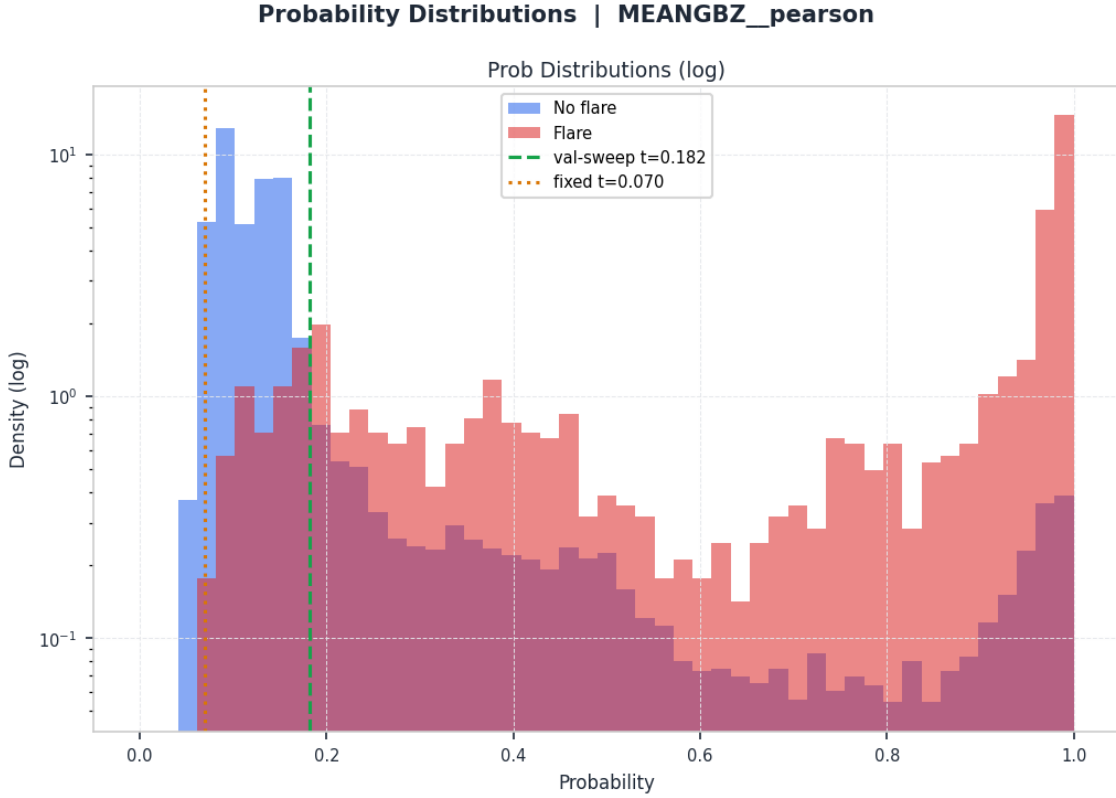


Figure 23: Output of the Pearson MEANGBZ model found at epoch 44 on the test samples. The vertical axis shows the sample density value on a log scale and the horizontal axis shows the probability score, or the model output. The red indicates a flare sample and the blue indicates a no flare sample, with dark red showing where they overlap. The swept threshold of the best run is shown as a dotted green line and the reference threshold is shown as the orange line.

The physics informed function seems to have pushed most samples up, and spread the no flare samples out over a larger range of probability scores. This cause the optimal threshold to be pushed up too, to avoid including too many false negatives. Looking a bit deeper into the model using the saliency analysis, depicted in 35, see appendix, for the history features, Figure 36, see appendix, for the delta features and solar phase cosine encoding and Figure 37, see appendix, for the SHARP features and solar phase sine encoding.

The saliency towards the later parts of the sequence is almost identical to the baseline model, apart from a few exceptions in the SHARP parameters, such as the increase for NPIX and decrease for MEANSHR. But overall the signal seems to be preserved, even through the physics informed loss functions influence. There is however a large difference in the beginning of the sequence, this makes sense as the gradient from the physics loss dominates in this area. What is however surprising is MEANGBZ not being the most salient feature even in the beginning where the physics informed loss

is most dominant. The most salient features are MEANGAM, MEANSHR and MEANGBT. Figure 38, see appendix, depicts a possible explanation for MEANGBT, as it also shows a large correlation alongside MEANGBZ. However, MEANSHR and MEANGAM on average show no correlation at all.

5.3.2 Worst model case study: Kendall MEANGAM

The confusion matrix for the Kendall MEANGAM model is shown in Table 10. While the number of false positives is relatively low compared to the baseline model, it has more than 2 times as much false negatives. This shows a clear difference to Pearson MEANGBZ model, which had a lower amount of false negative and a higher number of false positives, indicating a much higher tendency to classify samples as flares. The Kendall MEANGAM model misses quite a large number of flares.

Table 10: Confusion Matrix Kendall MEANGAM model at threshold 0.2173

	Predicted Negative	Predicted Positive
Actual Negative	TN = 28671	FP = 2962
Actual Positive	FN = 427	TP = 957

The worst performing model is Kendall MEANGAM, which had a vastly different training trajectory compared to the best model. The training and validation classification loss is displayed in Figure 24.

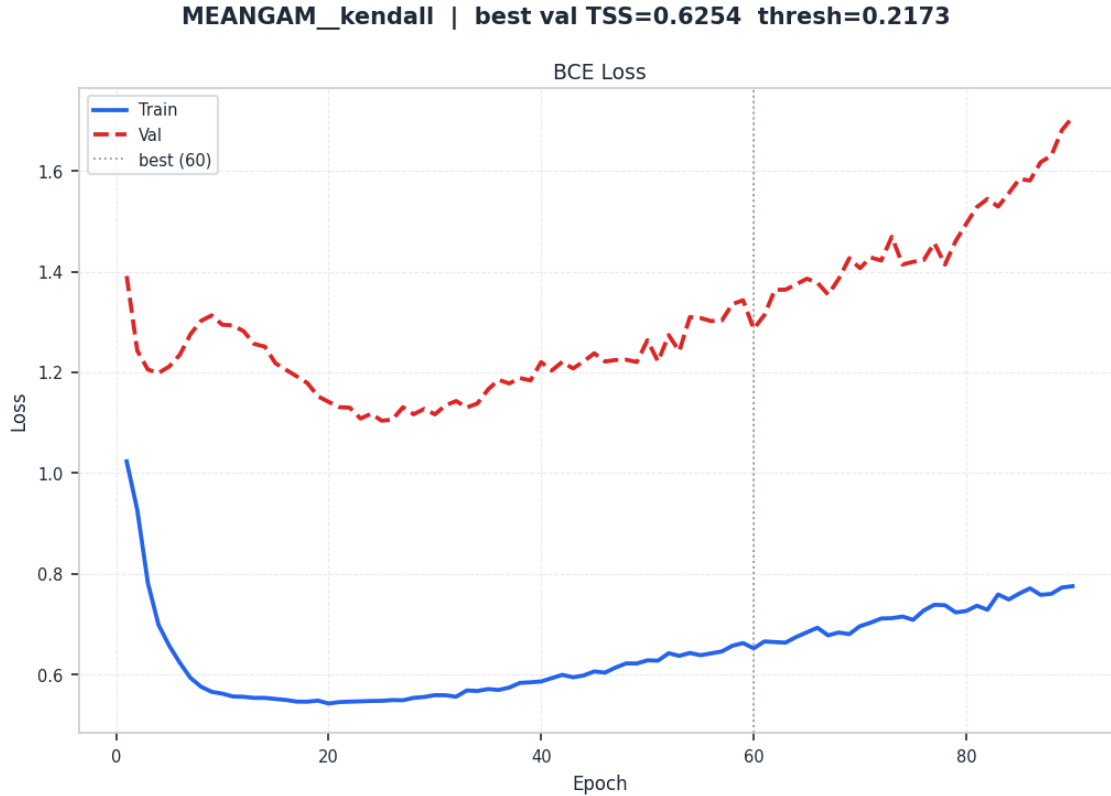


Figure 24: The vertical axis shows the value of the loss and the horizontal axis shows the corresponding epoch. The blue line indicates the training loss and the red line indicates the validation loss. The best epoch, epoch 60, is shown with a dotted gray line.

Compared to the previous model, the training loss is higher at every stage. With 0.6 at epoch 10, for an unscaled loss of 3 and 0.65 at epoch 50, for an unscaled loss of roughly 1.58. Meaning that this model also improves its classification of the training data, but performs worse than the best model. Additionally, the optimal score is found much later at epoch 60 compared to epoch 38. Looking at Figure 25, which depicts the TSS at each stage does however show an interesting contrast.

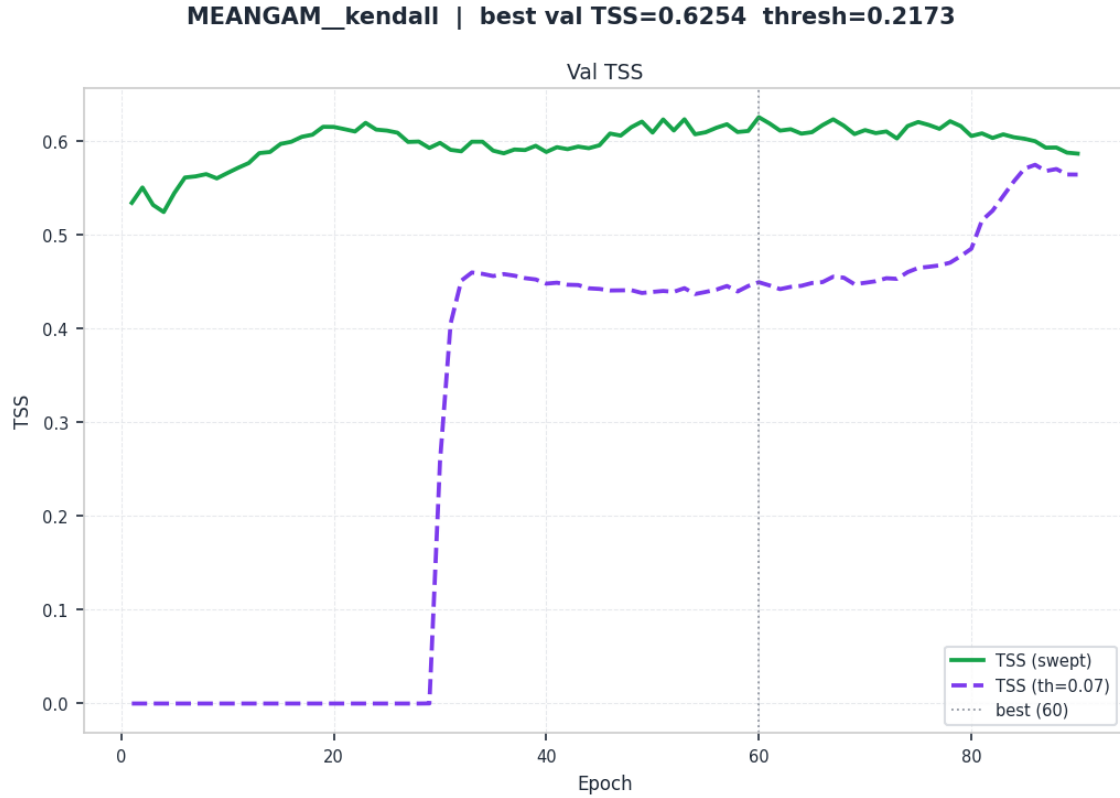


Figure 25: The vertical axis shows the validation TSS and the horizontal axis shows the corresponding epoch. The green line indicates the TSS corresponding to the sweeping threshold and the purple line indicates the TSS for the constant threshold. The best epoch, epoch 60, is shown with a dotted gray line.

The previous model shows a steady increase of the validation TSS, going from 0.1 at epoch 2, to a 0.7 at epoch 30. The Kendall MEANGAM model starts at a TSS of 0.55 at epoch 1 and slowly increases to 0.6 around epoch 20 and does not reach much higher from there. The physics loss, shown in Figure 26, is also higher.

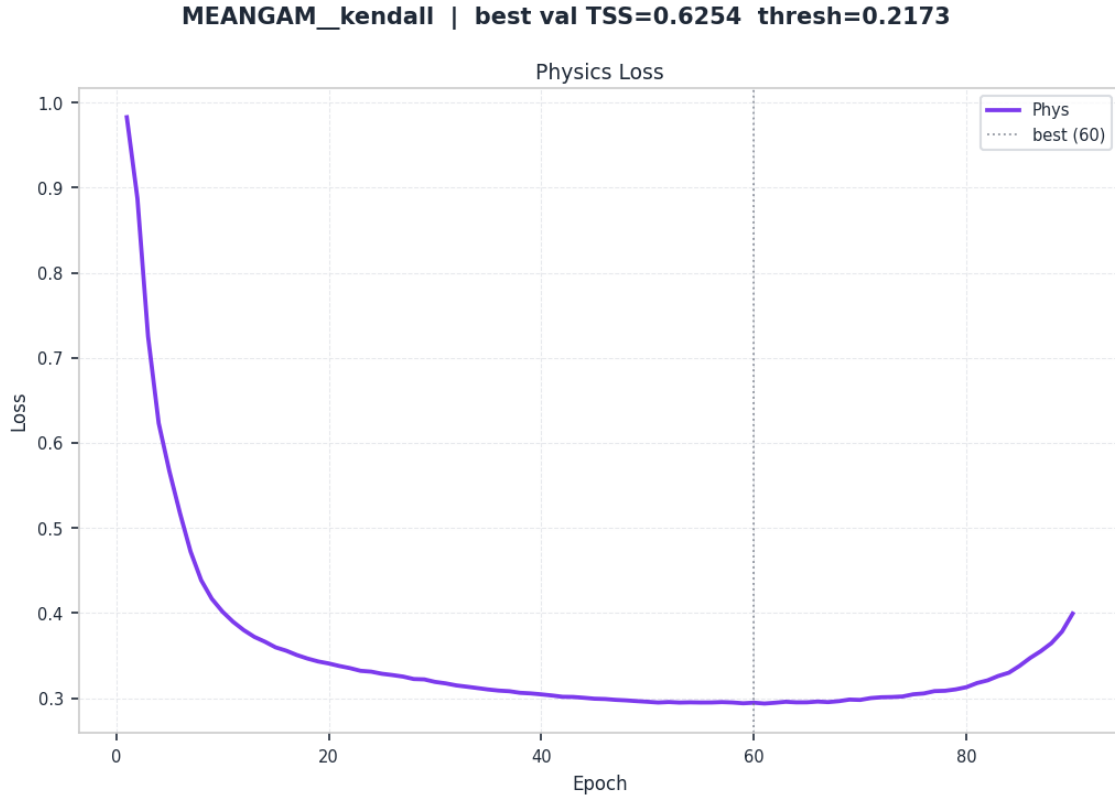


Figure 26: The vertical axis shows the value of the loss and the horizontal axis shows the corresponding epoch. The purple line indicates the physics loss. The best epoch, epoch 60, is shown with a dotted gray line.

However, because the Kendall MEANGAM model stops later, it does show something the other model does not. At roughly epoch 70, the physics loss starts to increase rapidly, demonstrating the model starting to heavily favor the classification loss over the physics loss, due to the scaling factor.

As for the threshold, the MEANGAM model finds an even higher optimal threshold at 0.2173. The distribution of the samples is depicted in Figure 27.

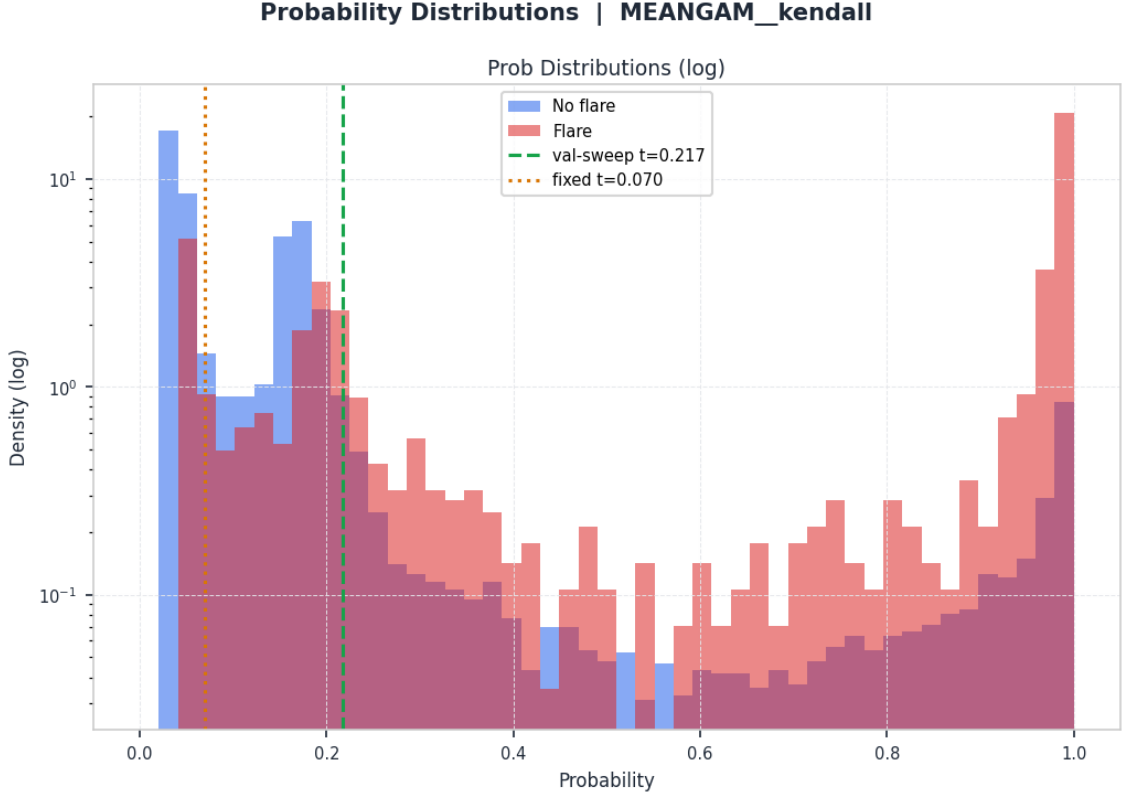


Figure 27: Output of the Kendall MEANGAM model found at epoch 60 on the test samples. The vertical axis shows the sample density value on a log scale and the horizontal axis shows the probability score, or the model output. The red indicates a flare sample and the blue indicates a no flare sample, with dark red showing where they overlap. The swept threshold of the best run is shown as a dotted green line and the reference threshold is shown as the orange line.

The Kendall MEANGAM model, much like the Pearson MEANGBZ model, shows a wider distribution of the no flare samples. A particularly large cluster is formed around probability 0.17 explaining the increased threshold to avoid these samples. This does however come at the cost of a larger amount of flare samples missed. The reason for the bad performance becomes clear when looking at the saliency analysis, depicted in 39, see appendix, for the history features, Figure 40, see appendix, for the delta features and solar phase cosine encoding and Figure 41, see appendix, for the SHARP features and solar phase sine encoding. The signal created by the classification loss towards the end of the sequence, which was present in the baseline model and the Pearson MEANGBZ model has largely disappeared in the Kendall MEANGAM model. What is again interesting is that despite being enforced in the physics loss function, MEANGAM is not the most salient feature, even in the beginning. Even though, as can be seen in Figure 42, it has the highest correlation to the model output, closely followed by both shear parameters.

5.4 Parameter study

As discussed in the methodology section, the parameter study consists of sequential elimination using permutation importance as a quantifier. The first 13 steps are depicted in Figure 28. The color indicates the change in TSS when a permutation was applied. In the first step the parameters Cdec, logEdec and MEANGBT are the most important. This is similar to what is shown in the saliency analysis of the baseline model, where Cdec and logEdec were the most important history features and MEANGBT was an important feature. However, some of the other high saliency features seem less important using this methodology. MEANGAM, MEANGBT and MEANSHR, which all had high saliency in the baseline model analysis, disappear among the other features in the permutation analysis. The first parameter that is dropped is the Bhis parameter, indicating that the occurrence of B-class flares does not give the model new information when there are this many parameters. At each step the model gets retrained on the parameters that are still used. As there are less and less parameters, each one becomes more useful. If a parameter was originally mostly redundant, it can become a valuable source of information when another parameter which provides similar information gets removed. In the second half, shown in Figure 29, it can be clearly seen that parameters become very important towards the latter steps. However, when looking at the evolution of the TSS score for each newly trained model, shown in Figures 43 and 44, see appendix, it can be seen that each retraining makes the model adapt to the new feature set a get back to roughly the same or an even higher TSS score. This is an important point that is not quite clear from looking at the importance score of features that are dropped. Which is depicted in Figure 45, see appendix. It can be seen that the first 8 features, change almost nothing or were even hurting the performance. At step 9 there is a slight turning point where each feature has a slight impact on the score, and at step 22 the first significant drop of at least 0.01 takes place. But even then, looking back at Figure 43 and Figure 44, when the model is retrained it seems to find the information it lost in the removed feature in another feature as the TSS remains stable. For the final test a set of features was selected based on the elimination process. The chosen features are: SolarPhase, TOTUSJH, MEANGBT, NPIX, ABSNJZH and Cdec. Where both solar encodings were used and each delta feature corresponding to the SHARP parameter for a total of 11 features. Representing an arbitrary selection of the best performing features from the elimination process, and a theoretical variety in feature type. While for example SIZE_ACR and NPIX are both present at the end, they represent very similar information, so SIZE_ACR is left out. The exact same model as the baseline was trained but instead of all, only the previously mentioned features were used as input. The validation TSS curve is shown in Figure 30. Interestingly, while there was much less input information it took actually much longer to find the optimal epoch, at epoch 81. The selected threshold was also higher at 0.2173, which suggests the model struggles with assigning extreme scores. The confusion matrix for the model applied to the test data is given in Table 11. With a TSS of 0.7152, which is really close to the score of the baseline model. This demonstrates that with a much smaller selection of features, the performance can still be achieved.

Table 11: Confusion Matrix baseline model with only 11 features at threshold 0.2173

	Predicted Negative	Predicted Positive
Actual Negative	TN = 27219	FP = 4414
Actual Positive	FN = 201	TP = 1183

Importance Heatmap — first 13 steps

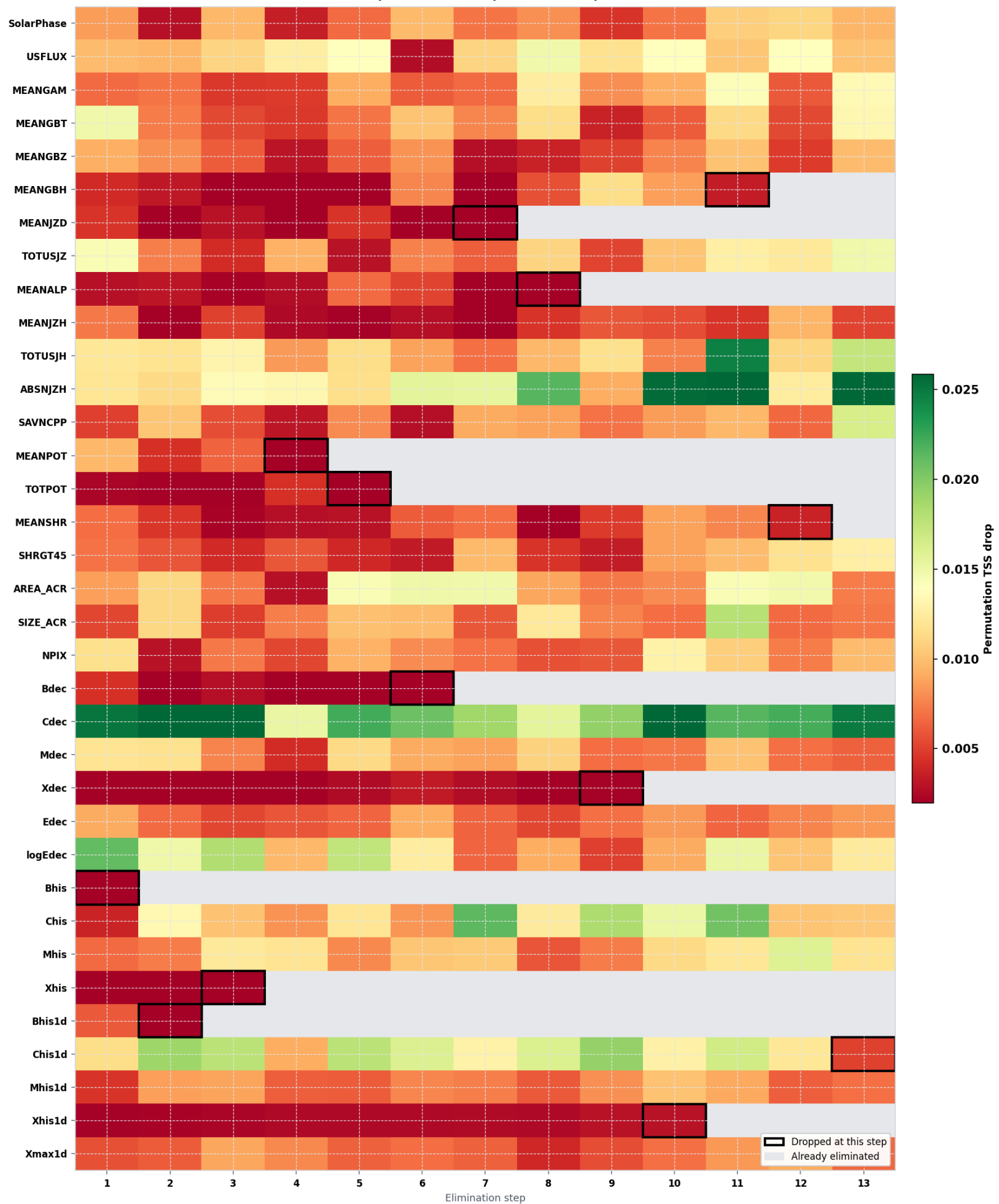


Figure 28: A heatmap of the TSS permutation per elimination step for each feature, for the first 13 steps.

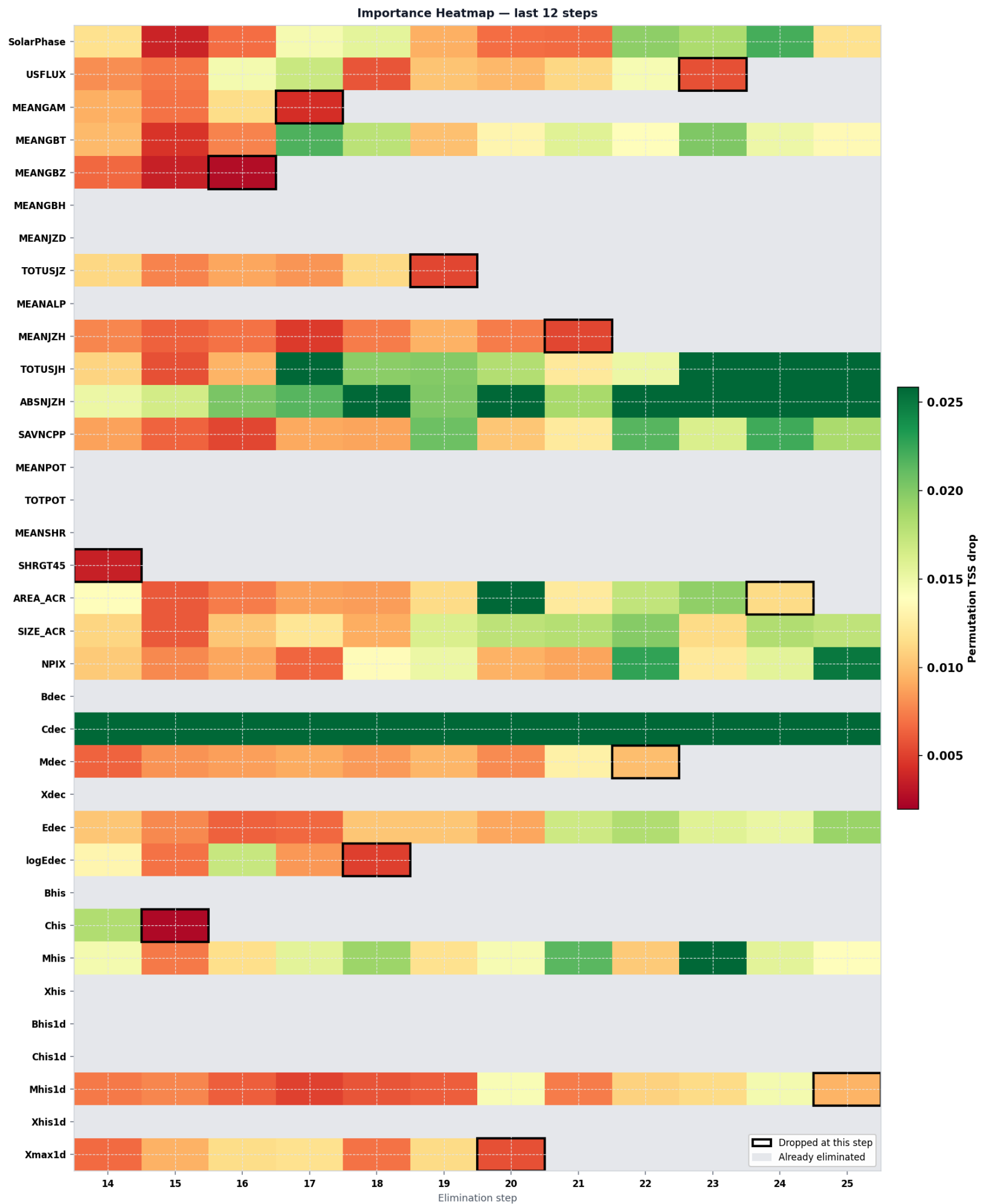


Figure 29: A heatmap of the TSS permutation per elimination step for each feature, for the last 12 steps.

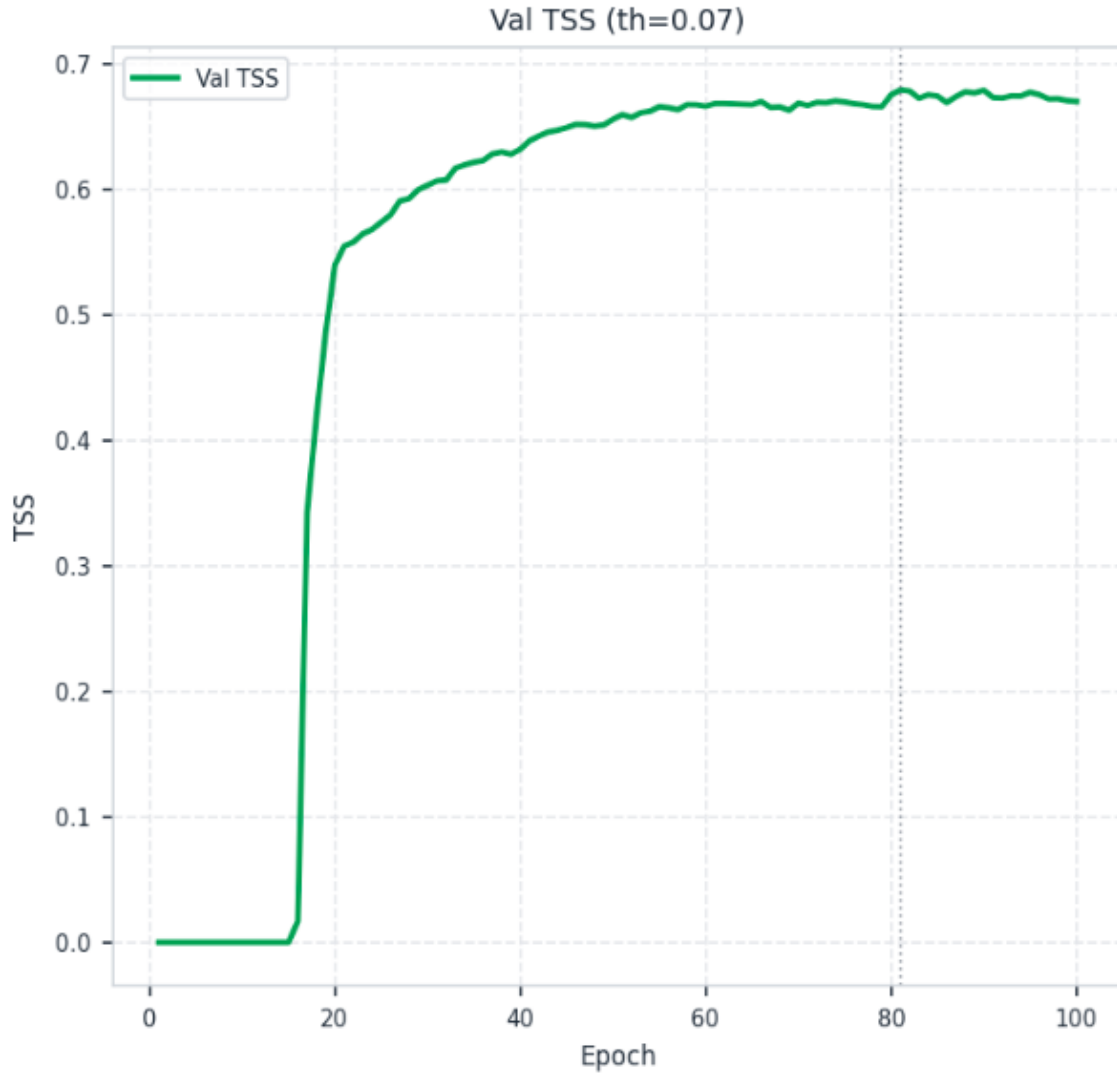


Figure 30: Validation TSS of the model using 11 parameters as input during training. The vertical axis represents the TSS score and horizontal axis the epoch number. The optimal TSS is achieved at epoch 81, indicated by the dotted gray line.

6 Discussion

6.1 Hyperparameter selection

In general, the tuning process showed that while most hyperparameters have an optimal value, there is a lot of room to increase or decrease the parameters without reducing the TSS score too much. There are a few reasons for this. First, the strongest aspect of the LSTM is the recurrent processing of information. This process is subject to a few parameters that were selected on a purely theoretical basis and were not tested for in this particular project. For example, sequence length, prediction window length, and timestep size are set at 24 hours, 24 hours, and 1 hour, respectively. Jiao et al. [61] find a similar TSS value of 0.74 for prediction of M and X class flares using a 24-hour prediction window. But setting the prediction model to 12 hours instead increases the TSS score to 0.88, and extending it to 36 hours results in a TSS of 0.70. Apart from the initial outlier in the search process, no configuration has this much effect on the TSS score.

As can be seen in Table 5, the actual value used for the positive weight is different from the value found in the best configuration. While the configuration that was found should theoretically result in the best TSS score, each parameter affects the model in different ways that may or may not be desirable.

First, the hidden size can be thought of as the model’s memory. While this means that a larger hidden size allows a model to remember more complex representations, it also means that the model becomes more prone to fitting noise, and the model becomes more computationally expensive. So, as the average performance of each hidden size is roughly equal, this parameter is kept small to reduce complexity and help combat overfitting. Similar arguments can be made for the number of layers. While more layers allow the model to form more complex representations of the data, they also increase the model’s complexity and computational load. But reducing this parameter to 1 would remove the dropout that occurs between layers 1 and 2 as well. Keeping the LSTM 2 layers allows for some more generalization through dropout in the sequential part of the model.

The dropout and weight decay both converged to quite high values. This indicates that the optimization process also finds that generalization is critical for this particular model. By having high weight decay and high dropout, the model is explicitly pushed towards finding general representations that do not strongly rely on 1 parameter. This is also seen in the low learning rate. A low learning rate was already the preference, as this allows for closer monitoring of the effects of 2 different loss functions. So, this low learning rate not only serves as a way to not have the model overfit but also as a way to make the model learn slower and more stable, which is important when the second loss function gets included.

Finally, the only parameter that has been changed from the found configuration is the positive weight. The positive weight is also the only value that has a theoretically ”correct” value. The class balance of the dataset is 26, meaning for each flare sample there are 26 non-flare samples. This would mean that the positive weight should theoretically be 26. However, it is set to 28 to slightly increase the value of flare samples, as missing a flare is much less desirable than mislabeling a non-flare sample.

6.2 Baseline model

The first notable thing about the baseline model is the presence of overfitting. The validation loss starts to rise around epoch 8 while the training loss keeps reducing after a little increase. The best model, based on validation TSS, is found at epoch 11. Interestingly, this is found through a sweeping threshold, which, through closer investigation, turns out to be a mismatch with the optimal testing threshold. The optimal validation threshold eventually converges with the reference threshold of 0.7 and the eventual optimal testing threshold turns out to be much closer to this threshold as well. This suggests that another methodology for finding a threshold might lead to better results. A possibility could be to use the plateau created by the sweeping threshold, which happens to be at the chosen reference threshold.

Looking at the TSS score, precision, and recall score compared to other similar models in Table 6.2. The TSS that was found is quite reasonable. While TSS is a specifically useful metric for comparing different models as it is invariant to class balance, using it directly to compare different models should be done with caution. For example, while Sun et al. [86] report an impressive TSS of 0.83, other factors can play a significant role. For example, instead of using a limb filter of 70° , a filter of 45° was used. In addition, the way in which the dataset for this model was constructed leads to a much healthier class balance of roughly 13% flare samples. Such effects can also be found in others and this one. For example, Abdullah et al. [87] found that when training and testing on the samples that all fall within the same year, different years have significantly varying scores. One interesting comparison is the use of a transformer layer, by Abdullah et al. [87] or an attention layer, by Liu et al. [71]. Both of these outperform the standard LSTMs present. A notable difference is that the LSTM by Abdullah et al. makes use of a sequence of length 10, cadence of 1 hour, and the LSTM from Liu et al. uses a sequence of length 12, also a cadence of 1 hour. Whereas the LSTM from this study and the LSTM of Jiao et al. both make use of a 24 hour sequence. Jiao et al. also report tests using a shorter sequence of 12 hours and a longer sequence of 48 hours, which both decrease the TSS by 0.02 and 0.01 respectively. This could suggest that the use of attention or transformer mechanisms is better at capturing long-term dependencies than extending the sequence length.

Table 12: Overview of comparable models also predicting M and X in a 24 window.

TSS	Recal	Precision	Model type	Reference
0.72	0.856	0.213	LSTM	Baseline
0.79	0.881	0.220	LSTM with attention	Liu et al. [71]
0.84	0.842	0.848	LSTM with transformer	Abdullah et al. [87]
0.70	-	-	LSTM	Abdullah et al. [87]
0.76	0.80	0.92	LSTM	Jiao et al. [61]
0.75	0.817	0.889	CNN	Li et al. [88]
0.68	0.730	0.282	LSTM	Wang et al. [89]
0.83	0.925	0.595	3D CNN	Sun et al. [86]

Moving on to the other results from the baseline model. The probability distributions show that the classes are not fully pushed to either extreme. This suggests that the model is quite conservative and does not make fully confident predictions. There are a few reasons for this. The most important reason is the use of early stopping. As the model gets trained, it gets more and more

confident about most samples; this pushes them to the extreme ends. However, as the model is stopped early, this stage is not quite reached. In addition, the selection of a high weight decay and a small learning rate limits the model’s ability to quickly make extreme judgments.

While the baseline model was not subject to any physics-based loss functions, it did have the head, which takes in the hidden state from each sample in the sequence. While the parameters inside this head are randomly initialized and not changed throughout the training, the hidden states do change based on the classification loss. It is interesting to see that some parameters do already show a correlation to the output used for the physics-based loss. The strongest correlation is the parameter TOTPOT, which is the total energy in a region.

From the saliency analysis, similar conclusions can be drawn compared to the model of Liu et al. [71]. The most important parameters are the SHARP features. These do differ from the important features found in the analysis by Liu et al., who found that TOTUSJH and TOTUSJZ, along with USFLUX and TOTPOT, are the most important parameters. Compared to the saliency analysis, which finds that MEANGAM and MEANGBT are the most important features, followed by MEANSHR, SHRGT45, MEANPOT, and SIZE_ACR. Where interestingly, SIZE_ACR seems to outperform AREA_ACR which corrects for projection effects. A possible explanation could be that leaving the distortion in place actually gives the model more information about the region’s characteristics. Additionally, the projection correction could introduce more noise to the parameter, weakening the signal. Another interesting difference is the low saliency of the TOTPOT parameter, found by Liu et al. [71] to be one of the most important ones and found to show the highest correlation with the output of the secondary model head. The most likely explanation for the low saliency of the TOTPOT is the high correlation between the parameters that causes the signal in TOTPOT to be redundant. This seems to be supported by the almost non-existent saliency of the delta feature of TOTPOT. Suggesting that the evolution of TOTPOT over time provides no real information to the model.

The final interesting parameters are the two solar phase encodings. These have not been used before in a solar flare prediction model but do show much promise. There is an interesting discrepancy between the cosine and sine encoding. Because the cosine varies most at the solar cycle peaks and valleys, it plays a more prevalent role in the prediction model. However, it should be noted that because of the total lack of any positive samples in the years 2018, 2019, and 2020, the solar phase parameter may appear more useful compared to a less extreme cycle. Nevertheless, the saliency analysis indicates that the model can derive useful information from this solar phase encoding.

Moving to the next set of figures. While the classes show good separation, the large class imbalance shows that threshold selection is highly dependent on what evaluation metric is used. If the F1 score were selected, a much higher threshold would be better. However, because TSS is chosen to best reflect the model performance, the threshold is pushed down to include as many true positives as possible. This highlights the importance of selecting a good evaluation metric based on the goal of the model. In addition, there is a large difference between the threshold found by the validation sweep and the optimal threshold based on test set results. This becomes a problem if the model is used in an operational context. A possible way to deal with this would be to keep adapting the threshold based on what results are achieved during operation to optimize performance.

The operational deployment test on data from the first 5 months of 2026 highlights the difficulties when moving from the trained model to operational deployment. While the model still performs on a decent level, with a TSS of 0.69 when using the optimal threshold. This optimal threshold is also different from both the optimal threshold for the test set and the validation set, showing again the importance of a proper threshold selection. For the TSS there is a clear decrease of 0.04. This is not unexpected as the model moves from a test set that follows the same distribution as the training set to a completely unknown distribution. It should also be noted that this set is rather small compared to the test set, so the performance is sensitive to a few bad predictions. On the other hand, this new set is still relatively close to the training data, and generally performance degrades for prediction models as time goes on due to changing circumstances. This is, of course, under the assumption that the model does not get trained or tuned on the new data.

6.3 Physics-informed model

The physics-informed models show a number of important differences compared to the baseline model, particularly in their training behavior, threshold selection, output distributions, and the balance between false positives and false negatives. The results also demonstrate that the choice of physical relationship used in the physics-informed loss has a substantial effect on the resulting model. The Pearson MEANGBZ model performs relatively close to the baseline, while the Kendall MEANGAM model performs considerably worse. The confusion matrices provide an additional perspective on these differences, showing that models with similar TSS values can have substantially different error profiles.

Starting with the best-performing physics-informed model, Pearson MEANGBZ, the classification loss shows a different behavior from that of the baseline model. The training loss is initially considerably lower, reaching approximately 0.4 at epoch 10 compared to approximately 0.9 for the baseline. However, this comparison is misleading because the classification loss is scaled by $(1 - \alpha)$. At epoch 10, $\alpha = 0.8$, meaning that the observed loss of 0.4 corresponds to an unscaled classification loss of approximately 2. Without the scaling, the classification loss would therefore actually be considerably higher than that of the baseline. By approximately epoch 50, α has decreased to 0.4, and the observed loss of approximately 0.5 corresponds to an unscaled loss of 1.25. This shows that the model is improving its classification of the training data throughout training, despite the apparent increase in the scaled classification loss. The best epoch, based on validation TSS, is found at epoch 38.

The physics loss provides an important additional perspective on this behavior. While the scaled physics loss remains relatively flat as the scaling factor changes, the corresponding unscaled physics loss increases. Comparing epoch 10, where $\alpha = 0.8$, with epoch 50, where $\alpha = 0.4$, the similar scaled physics loss therefore indicates that the actual physics loss has increased. This suggests that as training progresses, the model increasingly prioritizes the classification objective over the physical constraint. This behavior follows directly from the changing scaling factor: the classification loss receives an increasing weight, while the physics loss receives a decreasing weight. Consequently, later stages of training are increasingly focused on classification performance, even if this comes at the cost of satisfying the physical relationship.

The validation TSS of the Pearson MEANGBZ model is slightly lower than that of the baseline

model, although the difference is relatively small. When looking only at TSS, this could suggest that there is little practical difference between the two models. However, the confusion matrix reveals a more meaningful distinction. At the selected threshold of 0.1820, the Pearson MEANGBZ model produces 146 false negatives and 5,022 false positives, compared to 26,611 true negatives and 1,238 true positives. This means that the model misses substantially fewer flare samples than the baseline, but at the cost of incorrectly identifying more no-flare samples as flares.

This trade-off is particularly relevant for solar flare prediction, as the preferable model depends on the consequences associated with each type of error. A lower number of false negatives means that the Pearson MEANGBZ model is less likely to miss an actual flare, which may be preferable in an operational forecasting context where failing to warn about a flare can be particularly costly. On the other hand, the increase in false positives means that more false alarms would be generated. Therefore, the similar TSS between the baseline and Pearson MEANGBZ models hides a difference in their practical behavior. The physics-informed model has shifted the balance towards detecting more positive samples, rather than simply reproducing the behavior of the baseline.

This behaviour is also visible in the probability distribution of the Pearson MEANGBZ model. The physics-informed loss appears to push most samples towards higher probability scores while also spreading the no-flare samples over a larger range. As a consequence, the optimal threshold found through the validation sweep is considerably higher than the reference threshold. The increased threshold is necessary to prevent the wider distribution of no-flare samples from producing an excessive number of false positives. Despite this, the selected threshold still results in a relatively large number of false positives. This indicates that the physics-informed loss has changed the calibration of the model outputs rather than simply improving the separation between the two classes.

The higher threshold and altered probability distribution are therefore important when considering the model for practical use. The output probabilities of the physics-informed model cannot be interpreted in exactly the same way as those of the baseline model, and the reference threshold cannot simply be transferred between the models. The difference in the confusion matrices demonstrates why this is important: a relatively small change in overall TSS can correspond to a substantial change in the number of false positives and false negatives.

The saliency analysis provides further insight into how the Pearson MEANGBZ model manages to retain much of the classification performance of the baseline. Towards the later parts of the sequence, the saliency is almost identical to that of the baseline model, apart from a few changes in the SHARP parameters, such as an increased importance of NPIX and decreased importance of MEANSHR. This suggests that the useful classification signal learned by the baseline is largely preserved despite the addition of the physics-informed loss. The largest difference occurs towards the beginning of the sequence, where the gradient from the physics loss is strongest compared to the classification gradient.

Interestingly, MEANGBZ does not become the most salient feature, despite being the feature explicitly used in the physics-informed loss. Instead, MEANGAM, MEANSHR, and MEANGBT are among the most salient features. The high saliency of MEANGBT may be partially explained by its strong correlation with MEANGBZ. However, MEANSHR and MEANGAM do not show a similar average correlation with MEANGBZ, making their high saliency more difficult to explain. This

suggests that the physics-informed loss does not simply cause the model to rely more heavily on the feature used to construct the physical constraint. Instead, its influence may propagate through more complex relationships between the different input parameters.

The Kendall MEANGAM model shows a substantially different and less favorable behavior. Its training classification loss remains higher throughout training, with an observed loss of approximately 0.6 at epoch 10, corresponding to an unscaled loss of approximately 3, and approximately 0.65 at epoch 50, corresponding to an unscaled loss of roughly 1.58. The model therefore does improve its classification of the training data, but performs worse than the Pearson MEANGBZ model at every comparable stage. In addition, its best epoch is found much later, at epoch 60 compared to epoch 38 for Pearson MEANGBZ. This difference suggests that the choice of physical relationship has a substantial effect on how easily the classification objective can be learned.

The validation TSS further highlights the difference between the two models. The Pearson MEANGBZ model shows a steady improvement in validation TSS, reaching a relatively high value by approximately epoch 30. In contrast, the Kendall MEANGAM model starts at a relatively high TSS of approximately 0.55 at epoch 1, but only gradually increases to around 0.6 by epoch 20 and does not improve substantially afterwards. The model therefore fails to achieve the same level of performance as the Pearson MEANGBZ model. This suggests that the Kendall MEANGAM constraint interferes more strongly with the classification signal rather than providing a useful complementary source of information.

The confusion matrix makes this difference particularly clear. At its optimal threshold of 0.2173, the Kendall MEANGAM model produces 427 false negatives and 2,962 false positives, alongside 28,671 true negatives and 957 true positives. Compared to the Pearson MEANGBZ model, it therefore produces substantially fewer false positives but more than twice as many false negatives. This is almost the opposite behaviour of the Pearson MEANGBZ model. The Kendall MEANGAM model is consequently much more conservative in predicting flares, resulting in fewer false alarms but missing a considerably larger number of actual flares.

This difference is important because it shows why TSS alone does not fully describe the practical behavior of the models. The Pearson MEANGBZ and Kendall MEANGAM models differ strongly in the type of errors they make. Pearson MEANGBZ favours identifying more flare samples, resulting in fewer missed flares but more false alarms. Kendall MEANGAM does the opposite, producing fewer false alarms but failing to identify a much larger fraction of the actual flares. Depending on the intended application, either behavior could theoretically be preferred. However, for an operational flare forecasting system, the large number of false negatives produced by Kendall MEANGAM is particularly concerning because it means that a substantial number of actual flare events would go undetected.

The output distribution of the Kendall MEANGAM model is consistent with this behavior. Similar to Pearson MEANGBZ, the no-flare samples are spread over a relatively wide range of probability scores. However, Kendall MEANGAM has a particularly large concentration of no-flare samples around a probability of approximately 0.17. This explains why the optimal threshold is pushed upwards to 0.2173, a higher threshold is required to prevent these samples from being classified as flares. However, raising the threshold also removes genuine flare samples from the positive class,

explaining the large number of false negatives in the confusion matrix. The threshold therefore provides a direct link between the probability distribution and the model’s practical error profile.

The saliency analysis gives perhaps the clearest indication of the reason for the poor performance of the Kendall MEANGAM model. The classification signal towards the end of the sequence, which is present in the baseline model and largely preserved in the Pearson MEANGBZ model, has largely disappeared. This suggests that the Kendall MEANGAM physics-informed loss interferes more strongly with the temporal information that is useful for classification. It is again interesting that MEANGAM itself does not become the most salient feature, despite being directly enforced by the physics loss. This is consistent with the Pearson MEANGBZ model and suggests that directly including a feature in the physical constraint does not necessarily cause the model to use that feature more strongly for classification.

At the same time, MEANGAM has the highest correlation with the model output, closely followed by the two shear parameters. This highlights an important difference between correlation and saliency. A parameter can have a strong overall correlation with the model output without necessarily being the feature that contributes most strongly to the model’s individual predictions. In the Kendall MEANGAM model, the physical constraint appears to create a strong relationship between MEANGAM and the output, but this does not translate into a useful classification signal.

Overall, the results suggest that the effect of the physics-informed loss depends strongly on the physical relationship being imposed. The Pearson MEANGBZ model is able to retain much of the useful classification signal of the baseline and achieves a TSS only slightly below the baseline. More importantly, it changes the error balance in a potentially useful direction, reducing the number of missed flares from the baseline while accepting a larger number of false positives. The Kendall MEANGAM model, in contrast, produces fewer false positives but more than twice as many false negatives as Pearson MEANGBZ, indicating that the physical constraint interferes much more strongly with the classification task.

These results also demonstrate the limitations of evaluating the models using TSS alone. TSS is useful because it provides a class-balance-independent measure of performance, but two models with similar TSS can have very different confusion matrices and therefore very different operational characteristics. In this case, Pearson MEANGBZ and Kendall MEANGAM demonstrate almost opposing preferences in their classification errors. Examining precision, recall, false-positive rate, and false-negative rate alongside TSS is therefore necessary to understand the practical consequences of introducing the physics-informed loss.

Finally, the results raise questions about the current scaling strategy. The gradual reduction in the contribution of the physics loss means that the model is encouraged to prioritize classification more strongly as training progresses. While this can allow the model to improve its classification performance, it can also cause the physical constraint to become increasingly ignored, as demonstrated by the increasing physics loss in the later stages of Kendall MEANGAM training. Future work could therefore investigate alternative loss-weighting strategies that maintain a more consistent balance between the classification and physics objectives. The difference between the two models suggests that this balance is particularly important: a suitable physical constraint may be incorporated without substantially damaging the classification signal, whereas an unsuitable constraint or weighting

strategy can result in a model that misses a large number of actual flares.

Overall, the Pearson MEANGBZ model provides the strongest evidence that the physics-informed approach can be incorporated while maintaining performance close to the baseline. Its lower number of false negatives may even make it preferable to the baseline if detecting as many flares as possible is the primary objective. However, this comes with a substantial increase in false positives. The Kendall MEANGAM model demonstrates the opposite trade-off and, because of its much higher number of missed flares, is less attractive for an application where false negatives are particularly undesirable. Therefore, the main conclusion from these models is not that physics-informed learning universally improves performance, but that the choice and weighting of the physical constraint strongly determine both the overall performance and the type of errors the resulting model makes.

6.4 Parameter study

The parameter study provides an important indication that the full set of input parameters contains substantial redundancy. Rather than each parameter contributing unique information to the prediction, the sequential elimination process shows that the model can compensate for the removal of individual features by making use of information contained in the remaining parameters. This is particularly evident because the TSS of the retrained models remains relatively stable even when parameters that appear important according to the permutation analysis are removed.

The first notable result is the ranking of the parameters when all features are available. Cdec, logEdec, and MEANGBT are initially identified as the most important parameters. This agrees reasonably well with the saliency analysis of the baseline model, where Cdec and logEdec were the most salient history features and MEANGBT was also among the more important parameters. However, there are also some clear differences between the two approaches. MEANGAM, MEANGBT, and MEANSHR all showed relatively high saliency in the baseline analysis, whereas several of these parameters do not stand out to the same extent in the permutation analysis. This is not necessarily contradictory, as the two methods measure different concepts of importance. Saliency measures the sensitivity of the model output to changes in an input, whereas permutation importance measures how much the model’s performance changes when the information provided by a feature is disrupted.

The parameter study is particularly interesting when compared with the feature-selection approach used by Liu et al. [71]. Liu et al. also investigated whether all available input features were necessary and found that using only a subset of the most important features could outperform using all 40 features. Their LSTM models used 25 SHARP magnetic parameters together with 15 flare-history parameters, and they found that between 14 and 22 of the most important features were sufficient to achieve better performance than using the complete feature set. Their approach therefore provides a useful comparison for the current parameter study, as both studies arrive at the conclusion that reducing the input feature set does not necessarily harm performance.

There is, however, an important methodological difference. Liu et al. assessed feature importance in the context of their LSTM models and used the resulting ranking to investigate smaller feature subsets, whereas the parameter study in this work uses sequential elimination with permutation importance, retraining the model after every removal. The latter is particularly useful for investigating redundancy because the model is allowed to adapt to the reduced feature set after each

elimination. Consequently, the importance measured here should not be interpreted simply as a fixed ranking of which features are universally most informative. Instead, it measures how much the current model relies on each feature given the other features that are currently available.

This distinction can explain some of the differences between the results of this study and those of Liu et al. Liu et al. found that a relatively small subset of features could outperform the complete feature set, demonstrating that removing redundant or less useful features can improve the model. The present study finds a similar effect: the first eight removed features have almost no effect on performance and some even appear to slightly hurt the TSS. This suggests that these features either contain information that is already represented by other parameters or introduce information that does not generalize particularly well. The agreement between the two studies therefore strengthens the conclusion that using every available parameter is not necessarily optimal for solar flare prediction.

An important difference is that the current study also shows how feature importance changes as the feature set becomes smaller. The first parameter removed is Bhis, indicating that the occurrence of B-class flares provides little additional information when the model has access to the complete set of features. This does not mean that Bhis contains no information about future flaring. Instead, its information is likely already encoded by other history or magnetic parameters. Once those alternative sources of information are removed, however, the importance of the remaining features increases.

This effect becomes particularly clear in the second half of the elimination process. As fewer parameters remain, individual parameters become increasingly important according to permutation importance. At first glance, this could suggest that removing these parameters should cause increasingly large decreases in TSS. However, this is not what happens after retraining. The TSS curves of the newly trained models show that the model is generally able to recover the performance lost through the removal of a feature and sometimes even reaches a higher TSS. This demonstrates that the importance of a feature in the current model does not necessarily correspond to the amount of unique information that feature provides to a newly trained model.

This is an important distinction when interpreting feature importance. Suppose two parameters contain highly similar information. When both are available, permuting either one may have little effect because the model can rely on the other. After one parameter has been removed, the second may become highly important because it is now the primary source of that information. If the second parameter is then removed and the model is retrained, a third correlated feature may take over its role. The increasing importance of features towards the end of the elimination process is therefore likely to reflect the decreasing number of alternative information sources rather than an indication that those parameters suddenly become intrinsically more informative.

The evolution of the TSS supports this interpretation. The first eight eliminated features have little effect on the model's performance, while around step 9 there appears to be a slight turning point where the removal of individual features begins to have a more noticeable effect. The first significant reduction of at least 0.01 in TSS does not occur until step 22. Even at this point, however, the retrained model can largely recover the lost performance. This shows that the information removed from the model is not necessarily lost permanently. Instead, the remaining parameters can provide

alternative representations of the same underlying information.

The strong redundancy between SHARP parameters provides a plausible physical explanation for this behaviour. Many SHARP parameters describe related properties of the magnetic field and are strongly correlated with one another. For example, previous work has shown strong correlations between groups such as USFLUX, TOTUSJZ, TOTUSJH, and TOTPOT, while MEANPOT, SHRGT45, MEANSHR, and MEANGAM also form a highly correlated group. MEANGBZ and MEANGBT are particularly strongly correlated as well [90]. Therefore, it is not surprising that removing one parameter does not necessarily result in a large performance reduction. The model can often obtain similar information from another member of the same correlated group.

This also provides a possible explanation for the differences between the saliency analysis and the parameter study. For example, MEANGAM, MEANGBT, and MEANSHR can have high saliency because the model makes use of their information, but their permutation importance can be relatively low if another parameter provides similar information. In other words, a parameter can be highly useful to the model without being uniquely necessary. This distinction is particularly important for interpreting the results of the parameter study: the analysis is better understood as measuring the uniqueness of the information provided by a feature rather than simply its usefulness.

The comparison with Liu et al. is particularly relevant in this context. Their work used 40 features in total and found that models using only 14–22 of the most important features could achieve better performance than models using all 40. The current study reaches a similar conclusion using a somewhat different methodology. Rather than identifying a fixed optimal number of features solely from the initial ranking, the sequential elimination shows that the model can retain its performance through extensive feature removal because information can be redistributed among the remaining features. This suggests that feature reduction is not simply a matter of identifying the universally “best” parameters. Instead, several different feature combinations may provide similar predictive information.

For the final test, a reduced feature set was selected consisting of SolarPhase, TOTUSJH, MEANGBT, NPIX, ABSNJZH, and Cdec, together with both solar-phase encodings and the corresponding delta features for the SHARP parameters, resulting in a total of 11 features. This selection should not be interpreted as the uniquely optimal feature subset. Instead, it represents an arbitrary selection of strong-performing features from the elimination process while also attempting to retain variety in the types of information available to the model. For example, SIZE ACR and NPIX are both present towards the end of the elimination process and represent similar information, so SIZE ACR was excluded in favor of NPIX. This avoids unnecessarily including multiple parameters that provide largely overlapping information.

The resulting model performs remarkably close to the full baseline despite having considerably less input information. The test TSS is 0.7152, compared to approximately 0.72 for the full baseline model. The difference is therefore very small, demonstrating that most of the predictive performance can be retained with only 11 features. This result is consistent with the findings of Liu et al., who showed that reducing their 40 input features to a subset of 14–22 could maintain or even improve performance. The fact that two different feature-selection approaches lead to the same general conclusion provides stronger evidence that a large portion of the available feature set is

redundant.

The reduced model’s confusion matrix provides additional information about this small difference in TSS. At the selected threshold of 0.2173, the model produces 27,219 true negatives, 4,414 false positives, 201 false negatives, and 1,183 true positives. Compared to the full baseline, the reduced model therefore has fewer false positives but more false negatives. The model is consequently slightly more conservative in its flare predictions. This is an important distinction that is hidden when only comparing TSS: although the overall performance is almost unchanged, the reduced model has a somewhat different balance between detecting flares and generating false alarms.

The threshold itself is also noteworthy. The optimal threshold of 0.2173 is substantially higher than the reference threshold used for the baseline model. This suggests that the reduced model does not produce as many extreme probability scores and that a higher threshold is required to achieve its best TSS. This may be related to the reduced amount of information available to the model. With fewer features, the model may have less evidence with which to assign samples very high or very low probabilities, resulting in a more conservative probability distribution. However, this interpretation should be treated cautiously without directly comparing the output distributions of the two models.

Another interesting result is the considerably later optimal epoch. The reduced model reaches its best validation TSS at epoch 81, whereas the full baseline model reaches its optimum at epoch 11. Thus, reducing the input feature set does not make the model converge faster; in this case, the opposite occurs. One possible explanation is that the reduced model has to learn more complex relationships between the remaining features to recover information that was previously available directly through the removed parameters. This could lead to a slower optimization trajectory. Alternatively, the difference could simply arise from the stochastic nature of the training process and the altered optimization landscape produced by the smaller input space.

The results also suggest that removing features can potentially act as a form of regularization. The first eight eliminated parameters have almost no positive contribution to the final TSS, and some appear to slightly reduce performance. Removing such parameters may therefore reduce the amount of irrelevant or redundant information that the model has to process. This is consistent with Liu et al.’s finding that their reduced feature sets could outperform the full 40-feature model. It also provides a possible explanation for why the performance of the reduced model remains so close to that of the baseline despite its much smaller input space.

At the same time, the comparison with Liu et al. should not be interpreted as evidence that the exact same features are universally optimal. The two studies use different datasets, flare definitions, sequence lengths, model architectures, and feature-selection procedures. Liu et al. trained separate LSTM models for different flare classes using data from 2010 to 2018, with each sample containing 25 magnetic parameters and 15 flare-history parameters. Their model also used an attention mechanism and 10-hour sequences, whereas the model in this study uses a different architecture and feature representation. Consequently, agreement in the general finding of feature redundancy is more meaningful than agreement in the exact ranking of individual parameters.

The parameter study therefore provides an interesting contrast with the saliency analysis. Saliency suggests which features the existing model uses strongly, while sequential permutation elimination

reveals how much the model can compensate for the removal of those features. A feature can consequently have high saliency but low elimination importance if its information is easily replaced by another feature. Conversely, a feature can become highly important late in the elimination process because the alternatives that originally made it redundant have already been removed. Considering both analyses together gives a more complete picture of the model’s learned representation than either method alone.

Overall, the parameter study demonstrates that the full baseline feature set contains substantial redundancy and that the model does not require all available parameters to achieve good predictive performance. The first features removed have little influence on TSS, while later features become increasingly important as alternative information sources disappear. However, retraining allows the model to compensate for these removals, meaning that even relatively important features do not necessarily contain irreplaceable information. The final 11-feature model achieves a TSS of 0.7152, very close to the full baseline TSS of approximately 0.72, while using only a small fraction of the original inputs.

This result is also broadly consistent with Liu et al. [71], who found that reducing their 40-feature input to 14–22 important features could improve performance. Together, these findings suggest that feature reduction is a promising approach for solar flare prediction models. However, the current study additionally demonstrates that there may not be a single definitive ranking of features: their importance depends strongly on which other features are available. The most useful interpretation is therefore that the model contains groups of partially interchangeable information, allowing it to maintain almost the same predictive performance even after substantial feature removal. This is encouraging both for model interpretation and for practical deployment, as a smaller set of carefully selected features can provide nearly the same predictive skill while reducing the complexity of the model’s inputs.

6.5 Chapter overview

Most important discussion points:

1. Hyperparameter selection has little influence on performance. Theoretical arguments should be weighted against performance to make the correct decision.
2. Baseline model performance is relatively good, compared to others in literature. Attention mechanisms seem to slightly increase model performance. Deploying on temporally new data slightly decreases performance. Threshold selection matters a lot.
3. Physics informed loss function can be used to change underlying behavior of the model while maintaining performance. But to maintain performance, the model relies on the same signal as the baseline. So, the loss function can be used as a constraining tool but the underlying signal that the model finds should be similar.
4. Most parameters are highly correlated and can thus be removed while maintaining performance.

7 Verification

The discussion chapter identified several important findings regarding the physics-informed models and the redundancy of the input parameters. To verify these findings, four additional analyses were performed: an inspection of the individual hyperparameters and their theoretical interpretation, an evaluation using 2026 data, an analysis of the model probability distributions, and an independent test of the reduced 11-feature model.

7.1 Hyperparameter verification

The selected hyperparameters were verified by comparing the theoretically expected values with those obtained through the tuning process. In general, the results show that the tuned configuration is consistent with the theoretical considerations. The hidden size and number of layers were kept relatively small, which limits model complexity and reduces the risk of overfitting while still allowing the LSTM to learn meaningful temporal representations. Similarly, the relatively high dropout and weight decay values support the theoretical expectation that strong regularization is beneficial for this model. The low learning rate is also consistent with the expectation of slower and more stable learning, particularly when using the additional loss function.

The main deviation from the theoretically expected configuration is the positive weight. Based on the class imbalance of 26 non-flare samples for every flare sample, a positive weight of 26 would theoretically compensate for the imbalance. However, the final value of 28 was chosen to give flare samples slightly more importance, reflecting the higher cost of missing a flare. Overall, the comparison verifies that the final configuration is not simply the numerical optimum, but a balance between empirical performance, theoretical expectations, model complexity, and the practical importance of correctly detecting flares.

7.2 Verification using 2026 data

The models are additionally evaluated on data from 2026 to investigate whether the conclusions obtained from the original test period remain valid on later data. The models and previously selected thresholds are applied without further optimization, ensuring that 2026 acts as an independent verification period.

The performance of the model on the 2026 decreases slightly, compared to the test set. Showing that while still remain competitive it does seem to have a slight benefit from the data split being done based on active regions and not time.

7.3 Probability distribution verification

The probability distributions provide an additional explanation for the differences observed in the confusion matrices. Both physics-informed models produce a wider distribution of no-flare probabilities than the baseline, which explains why their optimal thresholds differ from the reference threshold.

For Pearson MEANGBZ, the higher threshold is consistent with the shift of no-flare samples towards higher probability values. Although increasing the threshold reduces false positives, a relatively large number remains. For Kendall MEANGAM, the concentration of no-flare samples around approximately 0.17 similarly explains why the optimal threshold increases to 0.2173. The higher threshold reduces false positives but also excludes more genuine flares, resulting in the larger number of false negatives.

This analysis therefore verifies that the differences in the confusion matrices are related to changes in the model output distributions. It also confirms that thresholds cannot simply be transferred between models. The physics-informed loss changes the calibration and distribution of the outputs, meaning that each model requires its own threshold.

7.4 Verification of the 11-feature model

The final verification evaluates the reduced model containing 11 features independently. The purpose is to determine whether the redundancy identified during the sequential elimination study remains evident after retraining the model using the reduced feature set.

The reduced model achieves a test TSS of 0.7152, compared with approximately 0.72 for the full baseline. Despite using substantially fewer inputs, the model therefore retains almost all of the predictive performance of the full model. This provides strong support for the conclusion that a considerable part of the original feature set is redundant.

The confusion matrix gives additional information. At a threshold of 0.2173, the reduced model produces 27,219 true negatives, 4,414 false positives, 201 false negatives and 1,183 true positives. Compared with the full baseline, the reduced model is somewhat more conservative, producing fewer false positives but more false negatives despite its similar TSS.

The reduced model also reaches its best validation performance considerably later, at approximately epoch 81 compared with epoch 11 for the baseline. This indicates that feature reduction does not necessarily simplify optimization. The model may instead need to learn more indirect relationships between the remaining parameters to recover information previously provided by the removed features.

8 Conclusion

The goal of this thesis was to improve the physical interpretability of solar flare prediction deep learning models. While this still remains a problem for deep learning models, progress was made, and based on the findings, an outline for creating solar flare prediction models and future research can be constructed. But first, the research questions should be answered.

Question 1: Which SHARP parameters and derived physical quantities provide the most relevant physical information for solar-flare forecasting?

Answer: The results indicate that physically relevant information is distributed across several groups of SHARP parameters rather than being concentrated in a single parameter. In particular, parameters associated with magnetic complexity, electric-current structure, magnetic shear, and magnetic-energy-related quantities provide important information for flare forecasting. The parameter study identified Cdec, logEdec, and MEANGBT as the most important parameters when all features were available. The saliency analysis of the baseline model additionally highlighted MEANGAM, MEANSHR, and MEANGBT, while the reduced-feature experiment showed that a model containing SolarPhase, TOTUSJH, MEANGBT, NPIX, ABSNJZH, and Cdec, together with their corresponding delta features and solar-phase encodings, could retain almost all of the predictive performance of the full model.

The results therefore support the hypothesis that physically motivated SHARP quantities contain substantial information about subsequent flare occurrence. However, they do not support the interpretation that there is one uniquely most informative SHARP parameter. Strong correlations between parameters such as USFLUX, TOTUSJZ, TOTUSJH, and TOTPOT, between MEANPOT, SHRGT45, MEANSHR, and MEANGAM, and particularly between MEANGBZ and MEANGBT, indicate substantial redundancy. Consequently, several different parameter combinations can provide similar predictive information. The parameter study demonstrates this clearly: the model retained a TSS of 0.7152 with only 11 features, compared with approximately 0.72 for the full baseline model.

Thus, the most relevant physical information is best understood in terms of groups of related quantities describing magnetic complexity, free magnetic energy, shear, and electric-current structure, rather than as a fixed ranking of individual SHARP parameters.

Question 2: How can established physical relationships between the selected parameters and solar-flare occurrence be incorporated into the loss function of a probabilistic deep-learning forecasting model?

Answer: The research showed that differentiable correlation-based loss terms can be used to encourage a relationship between the model’s learned representation and the cumulative evolution of a selected physical parameter. The constraint is therefore not applied directly between the physical parameter and the predicted flare probability. Instead, it acts on an intermediate representation of the input sequence, allowing the physical relationship to influence the learned representation while the classification loss continues to optimize the final forecasting output.

Question 2a: Which formulation of the physical loss function best represents the identified physical relationship while remaining compatible with the model’s probabilistic output?

Answer: Based on the results, the Pearson MEANGBZ formulation provides the strongest evidence for a useful physical constraint. It was able to impose the physical relationship while retaining much of the useful classification signal of the baseline model. The Pearson MEANGBZ model achieved a TSS comparable to the baseline and preserved much of the temporal classification information visible in the baseline saliency maps.

Question 2b: To what extent does constraining the learned representation using the physical loss function result in physically coherent model behaviour?

Answer: The results provide partial evidence that the physics-informed loss produces more physically constrained behaviour, but they do not demonstrate an unambiguous improvement in physical coherence across all formulations.

For the Pearson MEANGBZ model, the useful classification signal of the baseline was largely preserved, while the physical constraint altered the model’s probability distribution and error balance. The influence of the physical loss was strongest early in the sequence, where its gradient was relatively large compared with the classification gradient. Importantly, the model did not simply increase its reliance on MEANGBZ: other related parameters, particularly MEANGBT, MEANSHR, and MEANGAM, remained highly salient. This suggests that the physical constraint may propagate through the correlated structure of the SHARP parameters rather than producing a simple one-feature dependence.

The Kendall MEANGAM model provides a contrasting result. Although MEANGAM developed a strong correlation with the model output, this did not correspond to useful classification behaviour. Its late-sequence classification signal was substantially reduced, demonstrating that a strong statistical relationship between a physical parameter and the output does not necessarily imply physically useful or operationally coherent behavior.

Therefore, the results suggest that physical coherence depends strongly on the choice and weighting of the physical constraint. A suitable constraint can alter the model in a physically interpretable way without destroying the predictive representation, whereas an unsuitable constraint can dominate or interfere with information that is important for forecasting.

Question 3: What effect does the physics-informed loss function have on predictive performance compared with an otherwise identical data-driven model?

Answer: The results do not show a universal improvement in predictive performance. Instead, they demonstrate that the effect depends strongly on the physical relationship used in the loss.

The Pearson MEANGBZ model retained performance close to that of the baseline, with only a relatively small reduction in TSS. More importantly, it substantially changed the balance between false positives and false negatives. At its selected threshold of 0.1820, it produced 1,238 true positives, 146 false negatives, 26,611 true negatives, and 5,022 false positives. Thus, it missed considerably fewer flares but generated considerably more false alarms.

The Kendall MEANGAM model showed a substantially less favorable outcome. At its optimal threshold of 0.2173, it produced 957 true positives, 427 false negatives, 28,671 true negatives, and 2,962 false positives. It therefore adopted a more conservative forecasting behaviour, reducing false alarms at the expense of missing substantially more actual flares.

These results demonstrate that the physics-informed loss can change the operational characteristics of a forecasting model even when its TSS changes only modestly.

Question 3a: How does the physics-informed model compare with the standard model in terms of appropriate probabilistic solar-flare forecasting metrics?

Answers: TSS alone suggests that the Pearson MEANGBZ model performs similarly to the baseline, with only a relatively small difference. However, the confusion matrices reveal a much more substantial difference in practical behavior. The Pearson model has substantially fewer false negatives but substantially more false positives.

Consequently, TSS, precision, recall, false-positive rate, and false-negative rate should be considered together. The results specifically demonstrate why TSS alone is insufficient: two models with similar TSS can have very different operational consequences.

For an operational forecasting application in which missing an actual flare is particularly undesirable, the Pearson MEANGBZ model may therefore be preferable despite its increased false-alarm rate. Conversely, an application that places greater emphasis on avoiding false alarms could favor the more conservative behavior of the Kendall MEANGAM model, although its much larger number of missed flares is a significant disadvantage.

The physics-informed approach therefore maintained approximately baseline-level predictive skill for the Pearson formulation rather than clearly improving it, while substantially modifying the error profile. The Kendall formulation, by contrast, resulted in a clear degradation of the useful classification signal.

Question 3b: Does the predicted flare probability of the physics-informed model exhibit a stronger and more physically consistent relationship with the selected physical parameters than the standard model?

Answer: The physics-informed loss clearly changes the relationship between the model output and the physical parameters. In the Kendall MEANGAM model, for example, MEANGAM has the highest correlation with the model output, closely followed by the two shear parameters. This demonstrates that the physical constraint can create a strong statistical relationship between the selected physical parameter and predicted flare probability.

However, the saliency analysis shows that a strong correlation does not necessarily correspond to useful physical behaviour. MEANGAM's strong correlation in the Kendall model was not accompanied by a preservation of the useful classification signal. Similarly, MEANGBZ did not become the most salient feature in the Pearson MEANGBZ model despite being the parameter explicitly

used in the physical loss.

Therefore, the results suggest that the physics-informed loss does increase the influence of the imposed physical relationship on the model output, but they do not establish that this relationship is universally stronger or more physically meaningful than that of the baseline. Physical consistency must be evaluated using both output–parameter relationships and model-behavior measures such as saliency and temporal consistency.

8.1 Future outlook

Future work should focus on developing physics-informed solar-flare forecasting models that balance predictive accuracy, operational usefulness, and physical interpretability. The results show that the choice of physical relationship and its weighting in the loss function is critical. Future models should therefore investigate alternative loss-weighting strategies that maintain a meaningful physical constraint throughout training without suppressing the classification signal. In particular, physical relationships should be selected based not only on statistical correlation, but also on their physical relevance and compatibility with the forecasting task.

For operational applications, model evaluation should go beyond TSS and include precision, recall, false-positive and false-negative rates, calibration, and threshold-dependent performance. Since the physics-informed models produced substantially different probability distributions, their thresholds and probabilities should be calibrated independently rather than adopting those of a baseline model. The preferred model should ultimately depend on the operational cost of missed flares versus false alarms.

Interpretability should also be addressed using complementary methods. Saliency, feature importance, parameter correlations, and temporal analyses can reveal whether the model relies on physically meaningful information and whether the imposed physical constraint produces the intended behaviour. The strong redundancy between SHARP parameters further suggests that future models could use smaller, carefully selected feature sets, reducing complexity while retaining predictive performance.

Overall, the most promising direction is therefore not to impose as much physics as possible, but to identify well-supported physical relationships that complement the data-driven forecasting signal. Combining physically meaningful constraints, calibrated probabilistic outputs, robust operational metrics, and interpretable feature representations could lead to solar-flare forecasting systems that are both accurate and trustworthy for practical use.

References

- [1] Dabin Xue et al. “Examining the economic costs of the 2003 Halloween Storm effects on the North Hemisphere aviation using flight data in 2019”. In: *Space Weather* 21.3 (Mar. 2023). DOI: 10.1029/2022sw003381. URL: <https://doi.org/10.1029/2022sw003381>.
- [2] Tzu-Wei Fang et al. “Space weather environment during the SpaceX Starlink satellite loss in February 2022”. In: *Space Weather* 20.11 (Nov. 2022). DOI: 10.1029/2022sw003193. URL: <https://doi.org/10.1029/2022sw003193>.
- [3] en. URL: <https://www.swpc.noaa.gov/phenomena/solar-flares-radio-blackouts>.
- [4] L. Fletcher et al. “An observational overview of solar flares”. In: *Space Science Reviews* 159.1–4 (Aug. 2011), pp. 19–106. DOI: 10.1007/s11214-010-9701-8. URL: <https://doi.org/10.1007/s11214-010-9701-8>.
- [5] Scott Wiessinger and Karen Fox. *NASA’s SDO Observes a Cinco de Mayo Solar Flare*. en. May 2015. URL: <https://svs.gsfc.nasa.gov/11868/>.
- [6] Thomas N. Woods, Greg Kopp, and Phillip C. Chamberlin. “Contributions of the solar ultraviolet irradiance to the total solar irradiance during large flares”. In: *Journal of Geophysical Research Atmospheres* 111.A10 (Aug. 2006). DOI: 10.1029/2005ja011507. URL: <https://doi.org/10.1029/2005ja011507>.
- [7] B. T. Tsurutani et al. “The extreme magnetic storm of 1–2 September 1859”. In: *Journal of Geophysical Research Atmospheres* 108.A7 (July 2003). DOI: 10.1029/2002ja009504. URL: <https://doi.org/10.1029/2002ja009504>.
- [8] Norma B. Crosby, Markus J. Aschwanden, and Brian R. Dennis. “Frequency distributions and correlations of solar X-ray flare parameters”. In: *Solar Physics* 143.2 (Feb. 1993), pp. 275–299. DOI: 10.1007/bf00646488. URL: <https://doi.org/10.1007/bf00646488>.
- [9] Y. P. Li, W. Q. Gan, and L. Feng. “STATISTICAL ANALYSES ON THERMAL ASPECTS OF SOLAR FLARES”. In: *The Astrophysical Journal* 747.2 (Feb. 2012), p. 133. DOI: 10.1088/0004-637x/747/2/133. URL: <https://doi.org/10.1088/0004-637x/747/2/133>.
- [10] Jeffrey W. Reep and Kalman J. Knizhnik. “What determines the X-Ray intensity and duration of a solar flare?” In: *The Astrophysical Journal* 874.2 (Apr. 2019), p. 157. DOI: 10.3847/1538-4357/ab0ae7. URL: <https://doi.org/10.3847/1538-4357/ab0ae7>.
- [11] en. URL: <https://www.swpc.noaa.gov/noaa-scales-explanation>.
- [12] en. URL: https://www.spaceweather.gov/noaa-scales-explanation#r_scales.
- [13] Eelco Doornbos. *Thermospheric Density and Wind Determination from Satellite Dynamics*. Jan. 2012. DOI: 10.1007/978-3-642-25129-0. URL: <https://doi.org/10.1007/978-3-642-25129-0>.
- [14] J.T. Emmert. “Thermospheric mass density: A review”. In: *Advances in Space Research* 56.5 (June 2015), pp. 773–824. DOI: 10.1016/j.asr.2015.05.038. URL: <https://doi.org/10.1016/j.asr.2015.05.038>.
- [15] National Research Council. *Severe Space Weather Events—Understanding Societal and Economic Impacts*. Dec. 2008. DOI: 10.17226/12507. URL: <https://doi.org/10.17226/12507>.

- [16] D. H. Boteler. “A 21st century view of the March 1989 magnetic Storm”. In: *Space Weather* 17.10 (Oct. 2019), pp. 1427–1441. DOI: 10.1029/2019sw002278. URL: <https://doi.org/10.1029/2019sw002278>.
- [17] Arnold O. Benz. “Flare observations”. In: *Living Reviews in Solar Physics* 14.1 (Dec. 2016). DOI: 10.1007/s41116-016-0004-3. URL: <https://doi.org/10.1007/s41116-016-0004-3>.
- [18] E.R. Priest and T.G. Forbes. “The magnetic nature of solar flares”. In: *The Astronomy and Astrophysics Review* 10.4 (July 2001), pp. 313–377. DOI: 10.1007/s001590100013. URL: <https://doi.org/10.1007/s001590100013>.
- [19] Kazunari Shibata and Tetsuya Magara. “Solar flares: Magnetohydrodynamic processes”. In: *Living Reviews in Solar Physics* 8 (Jan. 2011). DOI: 10.12942/lrsp-2011-6. URL: <https://doi.org/10.12942/lrsp-2011-6>.
- [20] J Lin, W Soon, and S.L Baliunas. “Theories of solar eruptions: a review”. In: *New Astronomy Reviews* 47.2 (Mar. 2003), pp. 53–84. DOI: 10.1016/s1387-6473(02)00271-3. URL: [https://doi.org/10.1016/s1387-6473\(02\)00271-3](https://doi.org/10.1016/s1387-6473(02)00271-3).
- [21] Carolus J. Schrijver. “Driving major solar flares and eruptions: A review”. In: *Advances in Space Research* 43.5 (Nov. 2008), pp. 739–755. DOI: 10.1016/j.asr.2008.11.004. URL: <https://doi.org/10.1016/j.asr.2008.11.004>.
- [22] Chunming Zhu et al. “OBSERVATION OF THE EVOLUTION OF a CURRENT SHEET IN a SOLAR FLARE”. In: *The Astrophysical Journal Letters* 821.2 (Apr. 2016), p. L29. DOI: 10.3847/2041-8205/821/2/L29. URL: <https://doi.org/10.3847/2041-8205/821/2/L29>.
- [23] Sijie Yu et al. “Magnetic Reconnection during the Post-impulsive Phase of a Long-duration Solar Flare: Bidirectional Outflows as a Cause of Microwave and X-Ray Bursts”. In: *The Astrophysical Journal* 900.1 (Aug. 2020), p. 17. DOI: 10.3847/1538-4357/aba8a6. URL: <https://doi.org/10.3847/1538-4357/aba8a6>.
- [24] Eric Priest and Terry Forbes. *Magnetic reconnection*. May 2000. DOI: 10.1017/cbo9780511525087. URL: <https://doi.org/10.1017/cbo9780511525087>.
- [25] Ellen G. Zweibel and Masaaki Yamada. “Magnetic reconnection in astrophysical and laboratory plasmas”. In: *Annual Review of Astronomy and Astrophysics* 47.1 (Aug. 2009), pp. 291–332. DOI: 10.1146/annurev-astro-082708-101726. URL: <https://doi.org/10.1146/annurev-astro-082708-101726>.
- [26] J. F. Drake et al. “Magnetic reconnection in solar flares and the Near-Sun solar wind”. In: *Space Science Reviews* 221.2 (Mar. 2025). DOI: 10.1007/s11214-025-01153-x. URL: <https://doi.org/10.1007/s11214-025-01153-x>.
- [27] G. D. Holman et al. “Implications of x-ray observations for electron acceleration and propagation in solar flares”. In: *Space Science Reviews* 159.1–4 (Aug. 2011), pp. 107–166. DOI: 10.1007/s11214-010-9680-9. URL: <https://doi.org/10.1007/s11214-010-9680-9>.
- [28] Shin Toriumi and Haimin Wang. “Flare-productive active regions”. In: *Living Reviews in Solar Physics* 16.1 (May 2019), p. 3. DOI: 10.1007/s41116-019-0019-7. URL: <https://doi.org/10.1007/s41116-019-0019-7>.
- [29] Ryan French. “Where is the sun in its current 11-year solar cycle?” en. In: *Space* (Sept. 2025). URL: <https://www.space.com/astronomy/sun/where-are-we-in-the-current-11-year-solar-cycle>.

- [30] Phillip Chamberlin, William Pesnell, and Barbara Thompson. *The Solar Dynamics Observatory*. Jan. 2012. DOI: 10.1007/978-1-4614-3673-7. URL: <https://doi.org/10.1007/978-1-4614-3673-7>.
- [31] P. H. Scherrer et al. “The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO)”. In: *Solar Physics* 275.1–2 (Oct. 2011), pp. 207–227. DOI: 10.1007/s11207-011-9834-2. URL: <https://doi.org/10.1007/s11207-011-9834-2>.
- [32] URL: <https://122.physics.ucdavis.edu/?q=node/42>.
- [33] A. A. Norton et al. “Spectral line selection for HMI: a comparison of FE I 6173 Å and NI I 6768 Å”. In: *Solar Physics* 239.1–2 (Dec. 2006), pp. 69–91. DOI: 10.1007/s11207-006-0279-y. URL: <https://doi.org/10.1007/s11207-006-0279-y>.
- [34] J. M. Borrero et al. “VFISV: Very fast inversion of the Stokes vector for the helioseismic and magnetic imager”. In: *Solar Physics* 273.1 (Feb. 2010), pp. 267–293. DOI: 10.1007/s11207-010-9515-6. URL: <https://doi.org/10.1007/s11207-010-9515-6>.
- [35] S. Couvidat et al. “Observables processing for the helioseismic and magnetic Imager instrument on the Solar Dynamics Observatory”. In: *Solar Physics* 291.7 (Aug. 2016), pp. 1887–1938. DOI: 10.1007/s11207-016-0957-3. URL: <https://doi.org/10.1007/s11207-016-0957-3>.
- [36] M. G. Bobra et al. “The Helioseismic and Magnetic Imager (HMI) Vector Magnetic Field Pipeline: SHARPS – Space-Weather HMI Active Region patches”. In: *Solar Physics* 289.9 (Apr. 2014), pp. 3549–3578. DOI: 10.1007/s11207-014-0529-3. URL: <https://doi.org/10.1007/s11207-014-0529-3>.
- [37] K. D. Leka and G. Barnes. “Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. I. Data, General Approach, and Sample Results”. In: *The Astrophysical Journal* 595.2 (Sept. 2003), pp. 1277–1295. DOI: 10.1086/377511. URL: <https://doi.org/10.1086/377511>.
- [38] K. D. Leka and G. Barnes. “Photospheric Magnetic Field Properties of Flaring versus Flare-quiet Active Regions. IV. A Statistically Significant Sample”. In: *The Astrophysical Journal* 656.2 (Feb. 2007), pp. 1173–1186. DOI: 10.1086/510282. URL: <https://doi.org/10.1086/510282>.
- [39] Thomas N. Woods et al. “GOES-R Series X-Ray Sensor (XRS): 1. Design and Pre-Flight calibration”. In: *Journal of Geophysical Research Space Physics* 129.11 (Nov. 2024), e2024JA032925. DOI: 10.1029/2024ja032925. URL: <https://doi.org/10.1029/2024ja032925>.
- [40] Nicola Plutino et al. “A new catalogue of solar flare events from soft X-ray GOES signal in the period 1986–2020”. In: *Advances in Space Research* 71.4 (Nov. 2022), pp. 2048–2058. DOI: 10.1016/j.asr.2022.11.020. URL: <https://doi.org/10.1016/j.asr.2022.11.020>.
- [41] Mingfu Shao et al. “Advances and Challenges in Solar Flare Prediction: A review”. In: *arXiv (Cornell University)* (Nov. 2025). DOI: 10.48550/arxiv.2511.20465. URL: <https://doi.org/10.48550/arxiv.2511.20465>.
- [42] Zheng Sun et al. “Magnetic helicity evolution during active region emergence and subsequent flare productivity”. In: *arXiv (Cornell University)* (Mar. 2024). DOI: 10.48550/arxiv.2403.18354. URL: <http://arxiv.org/abs/2403.18354>.

- [43] Per Bak, Chao Tang, and Kurt Wiesenfeld. “Self-organized criticality: An explanation of the $1/f$ noise”. In: *Phys. Rev. Lett.* 59 (4 July 1987), pp. 381–384. DOI: 10.1103/PhysRevLett.59.381. URL: <https://link.aps.org/doi/10.1103/PhysRevLett.59.381>.
- [44] Edward T. Lu and Russell J. Hamilton. “Avalanches and the distribution of solar flares”. In: *The Astrophysical Journal* 380 (Oct. 1991), p. L89. DOI: 10.1086/186180. URL: <https://doi.org/10.1086/186180>.
- [45] Leonidas Karakatsanis and Georgios Pavlos. “SOC and Chaos into the Solar Activity”. In: *Nonlinear Phenomena in Complex Systems* 11 (Jan. 2008), pp. 280–284.
- [46] Ting Li et al. “Survey of Magnetic Field Parameters Associated with Large Solar Flares”. In: *The Astrophysical Journal* 964.2 (Mar. 2024), p. 159. DOI: 10.3847/1538-4357/ad2e90. URL: <https://doi.org/10.3847/1538-4357/ad2e90>.
- [47] Kanya Kusano et al. “A physics-based method that can predict imminent large solar flares”. In: *Science* 369.6503 (2020), pp. 587–591. DOI: 10.1126/science.aaz2511. eprint: <https://www.science.org/doi/pdf/10.1126/science.aaz2511>. URL: <https://www.science.org/doi/abs/10.1126/science.aaz2511>.
- [48] R. G. Giovanelli. “The relations between eruptions and sunspots.” In: *The Astrophysical Journal* 89 (June 1939), p. 555. DOI: 10.1086/144081. URL: <https://doi.org/10.1086/144081>.
- [49] Patrick S. McIntosh. “The classification of sunspot groups”. In: *Solar Physics* 125.2 (Jan. 1990), pp. 251–267. DOI: 10.1007/bf00158405. URL: <https://doi.org/10.1007/bf00158405>.
- [50] D. Shaun Bloomfield et al. “TOWARD RELIABLE BENCHMARKING OF SOLAR FLARE FORECASTING METHODS”. In: *The Astrophysical Journal Letters* 747.2 (Feb. 2012), p. L41. DOI: 10.1088/2041-8205/747/2/L41. URL: <https://doi.org/10.1088/2041-8205/747/2/L41>.
- [51] Kangjin Lee et al. “Solar Flare Occurrence Rate and Probability in Terms of the Sunspot Classification Supplemented with Sunspot Area and Its Changes”. In: *Solar Physics* 281.2 (Sept. 2012), pp. 639–650. DOI: 10.1007/s11207-012-0091-9. URL: <https://doi.org/10.1007/s11207-012-0091-9>.
- [52] Manolis K. Georgoulis and David M. Rust. “Quantitative Forecasting of Major Solar Flares”. In: *The Astrophysical Journal* 661.1 (Apr. 2007), p. L109. DOI: 10.1086/518718. URL: <https://doi.org/10.1086/518718>.
- [53] Hui Song et al. “Statistical assessment of photospheric magnetic features in imminent solar flare predictions”. In: *Solar Physics* 254.1 (Nov. 2008), pp. 101–125. DOI: 10.1007/s11207-008-9288-3. URL: <https://doi.org/10.1007/s11207-008-9288-3>.
- [54] N. Nishizuka et al. “Solar Flare Prediction Model with Three Machine-learning Algorithms using Ultraviolet Brightening and Vector Magnetograms”. In: *The Astrophysical Journal* 835.2 (Jan. 2017), p. 156. DOI: 10.3847/1538-4357/835/2/156. URL: <https://doi.org/10.3847/1538-4357/835/2/156>.
- [55] Julia Bringewald. “Solar Flare Forecast: A comparative analysis of machine learning algorithms for solar flare class prediction”. In: *arXiv (Cornell University)* (May 2025). DOI: 10.48550/arxiv.2505.03385. URL: <http://arxiv.org/abs/2505.03385>.

- [56] Griffin T. Goodwin, Viacheslav M. Sadykov, and Petrus C. Martens. “Investigating performance trends of simulated real-time solar flare predictions: the impacts of training windows, data volumes, and the solar cycle”. In: *The Astrophysical Journal* 964.2 (Mar. 2024), p. 163. DOI: 10.3847/1538-4357/ad276c. URL: <https://doi.org/10.3847/1538-4357/ad276c>.
- [57] Andrés Asensio Ramos et al. “Machine learning in solar physics”. In: *Living Reviews in Solar Physics* 20.1 (July 2023), p. 4. ISSN: 1614-4961. DOI: 10.1007/s41116-023-00038-x. URL: <https://doi.org/10.1007/s41116-023-00038-x>.
- [58] N. Nishizuka et al. “Deep Flare Net (DEFN) model for solar flare prediction”. In: *The Astrophysical Journal* 858.2 (May 2018), p. 113. DOI: 10.3847/1538-4357/aab9a7. URL: <https://doi.org/10.3847/1538-4357/aab9a7>.
- [59] Naoto Nishizuka et al. “Operational solar flare prediction model using Deep Flare Net”. In: *arXiv (Cornell University)* (Dec. 2021). DOI: 10.48550/arxiv.2112.00977. URL: <https://doi.org/10.48550/arxiv.2112.00977>.
- [60] Xin Huang et al. “Deep Learning based Solar Flare Forecasting Model. i. Results for line-of-sight magnetograms”. In: *The Astrophysical Journal* 856.1 (Mar. 2018), p. 7. DOI: 10.3847/1538-4357/aaae00. URL: <https://doi.org/10.3847/1538-4357/aaae00>.
- [61] Zhenbang Jiao et al. “Solar flare intensity prediction with machine learning models”. In: *Space Weather* 18.7 (May 2020). DOI: 10.1029/2020sw002440. URL: <https://doi.org/10.1029/2020sw002440>.
- [62] Rongxin Tang et al. “Solar flare prediction based on the fusion of multiple deep-learning models”. In: *The Astrophysical Journal Supplement Series* 257.2 (Dec. 2021), p. 50. DOI: 10.3847/1538-4365/ac249e. URL: <https://doi.org/10.3847/1538-4365/ac249e>.
- [63] Ashish Vaswani et al. *Attention is all you need*. en. June 2017. URL: <https://arxiv.org/abs/1706.03762>.
- [64] Kanta Kaneda et al. “Flare Transformer: Solar flare prediction using magnetograms and Sunspot physical features”. In: *Lecture Notes in Computer Science* (Jan. 2023), pp. 442–457. DOI: 10.1007/978-3-031-26284-5_27. URL: https://doi.org/10.1007/978-3-031-26284-5_27.
- [65] Yasser Abdullah et al. “Operational prediction of solar flares using a transformer-based framework”. In: *Scientific Reports* 13.1 (Aug. 2023), p. 13665. DOI: 10.1038/s41598-023-40884-1. URL: <https://doi.org/10.1038/s41598-023-40884-1>.
- [66] Liang Qiao and Gang Qin. “Flare Set-Prediction Transformer: A Transformer-Based Set-Prediction model for detailed solar flare forecasting”. In: *Universe* 11.6 (May 2025), p. 174. DOI: 10.3390/universe11060174. URL: <https://doi.org/10.3390/universe11060174>.
- [67] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215. DOI: 10.1038/s42256-019-0048-x. URL: <https://www.nature.com/articles/s42256-019-0048-x>.
- [68] M. Raissi, P. Perdikaris, and G.E. Karniadakis. “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”. In: *Journal of Computational Physics* 378 (2019), pp. 686–707. ISSN: 0021-9991. DOI: <https://doi.org/10.1016/j.jcp.2018.10.045>. URL: <https://www.sciencedirect.com/science/article/pii/S0021999118307125>.

- [69] Ben Moseley, Andrew Markham, and Tarje Nissen-Meyer. “Solving the wave equation with physics-informed deep learning”. In: *arXiv (Cornell University)* (June 2020). DOI: 10.48550/arxiv.2006.11894. URL: <http://arxiv.org/abs/2006.11894>.
- [70] M. G. Bobra and S. Couvidat. “SOLAR FLARE PREDICTION USING SDO/HMI VECTOR MAGNETIC FIELD DATA WITH A MACHINE-LEARNING ALGORITHM”. In: *The Astrophysical Journal* 798.2 (Jan. 2015), p. 135. DOI: 10.1088/0004-637x/798/2/135. URL: <https://doi.org/10.1088/0004-637x/798/2/135>.
- [71] Hao Liu et al. “Predicting solar flares using a long short-term memory network”. In: *The Astrophysical Journal* 877.2 (June 2019), p. 121. DOI: 10.3847/1538-4357/ab1b3c. URL: <https://doi.org/10.3847/1538-4357/ab1b3c>.
- [72] Eric Jonas et al. “Flare prediction using photospheric and coronal image data”. In: *Solar Physics* 293.3 (Feb. 2018). DOI: 10.1007/s11207-018-1258-9. URL: <https://doi.org/10.1007/s11207-018-1258-9>.
- [73] Yanfang Zheng et al. “Comparative analysis of machine learning models for solar flare prediction”. In: *Astrophysics and Space Science* 368.7 (July 2023). DOI: 10.1007/s10509-023-04209-y. URL: <https://doi.org/10.1007/s10509-023-04209-y>.
- [74] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
- [75] dida Datenshmiede GmbH. *What is an LSTM Neural Network?* en. URL: <https://dida.do/what-is-an-lstm-neural-network>.
- [76] T. N. Bungey and T. G. Forbes. “On current sheet approximations in models of eruptive flares”. In: *Solar Physics* 149.1 (Jan. 1994), pp. 205–208. DOI: 10.1007/bf00645190. URL: <https://doi.org/10.1007/bf00645190>.
- [77] Masaaki Yamada, Russell Kulsrud, and Hantao Ji. “Magnetic reconnection”. In: *Reviews of Modern Physics* 82.1 (Mar. 2010), pp. 603–664. DOI: 10.1103/revmodphys.82.603. URL: <https://doi.org/10.1103/revmodphys.82.603>.
- [78] J. W. Dungey. “The motion of magnetic fields”. In: *Monthly Notices of the Royal Astronomical Society* 113.6 (Dec. 1953), pp. 679–682. DOI: 10.1093/mnras/113.6.679. URL: <https://doi.org/10.1093/mnras/113.6.679>.
- [79] H. Alfvén. “Existence of Electromagnetic-Hydrodynamic waves”. In: *Nature* 150.3805 (Oct. 1942), pp. 405–406. DOI: 10.1038/150405d0. URL: <https://doi.org/10.1038/150405d0>.
- [80] Chengcai Shen et al. “The origin of underdense plasma downflows associated with magnetic reconnection in solar flares”. In: *Nature Astronomy* 6.3 (Jan. 2022), pp. 317–324. DOI: 10.1038/s41550-021-01570-2. URL: <https://doi.org/10.1038/s41550-021-01570-2>.
- [81] Xiaoli Yan et al. “Fast plasmoid-mediated reconnection in a solar flare”. In: *Nature Communications* 13.1 (Feb. 2022), p. 640. DOI: 10.1038/s41467-022-28269-w. URL: <https://doi.org/10.1038/s41467-022-28269-w>.
- [82] Takuya Akiba et al. “Optuna: A Next-generation Hyperparameter Optimization Framework”. In: *CoRR* abs/1907.10902 (2019). arXiv: 1907.10902. URL: <http://arxiv.org/abs/1907.10902>.
- [83] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *J. Mach. Learn. Res.* 3.null (Mar. 2003), pp. 1157–1182. ISSN: 1532-4435.

- [84] Gavin C. Cawley and Nicola L. C. Talbot. “On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation”. In: *Journal of Machine Learning Research* 11.70 (2010), pp. 2079–2107. URL: <http://jmlr.org/papers/v11/cawley10a.html>.
- [85] Christoph Molnar et al. “Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach”. In: *Data Mining and Knowledge Discovery* 38.5 (Jan. 2023), pp. 2903–2941. DOI: 10.1007/s10618-022-00901-9. URL: <https://doi.org/10.1007/s10618-022-00901-9>.
- [86] Pengchao Sun et al. “Solar flare forecast using 3D convolutional neural networks”. In: *The Astrophysical Journal* 941.1 (Dec. 2022), p. 1. DOI: 10.3847/1538-4357/ac9e53. URL: <https://doi.org/10.3847/1538-4357/ac9e53>.
- [87] René Prédal. “1. Le free cinéma britannique: Karel Reisz - Tony Richardson - Lindsay Anderson”. In: *CinémAction* 13.142 (Jan. 2012), pp. 146–13665. DOI: 10.1038/s41598-023-40884-1. URL: <https://dialnet.unirioja.es/servlet/articulo?codigo=3903012>.
- [88] Xuebao Li et al. “Predicting solar flares using a novel deep convolutional neural network”. In: *The Astrophysical Journal* 891.1 (Feb. 2020), p. 10. DOI: 10.3847/1538-4357/ab6d04. URL: <https://doi.org/10.3847/1538-4357/ab6d04>.
- [89] Xiantong Wang et al. “Predicting Solar Flares with Machine Learning: Investigating Solar Cycle Dependence”. In: *The Astrophysical Journal* 895.1 (May 2020), p. 3. DOI: 10.3847/1538-4357/ab89ac. URL: <https://doi.org/10.3847/1538-4357/ab89ac>.
- [90] Yang Chen et al. “Identifying solar flare precursors using time series of SDO/HMI images and SHARP parameters”. In: *Space Weather* 17.10 (Aug. 2019), pp. 1404–1426. DOI: 10.1029/2019sw002214. URL: <https://doi.org/10.1029/2019sw002214>.

9 Appendix

Module 3 -- Physics Correlation Analysis | baseline

SHARP Feature <-> Seq-Prob Pearson Correlation (train/val/test)

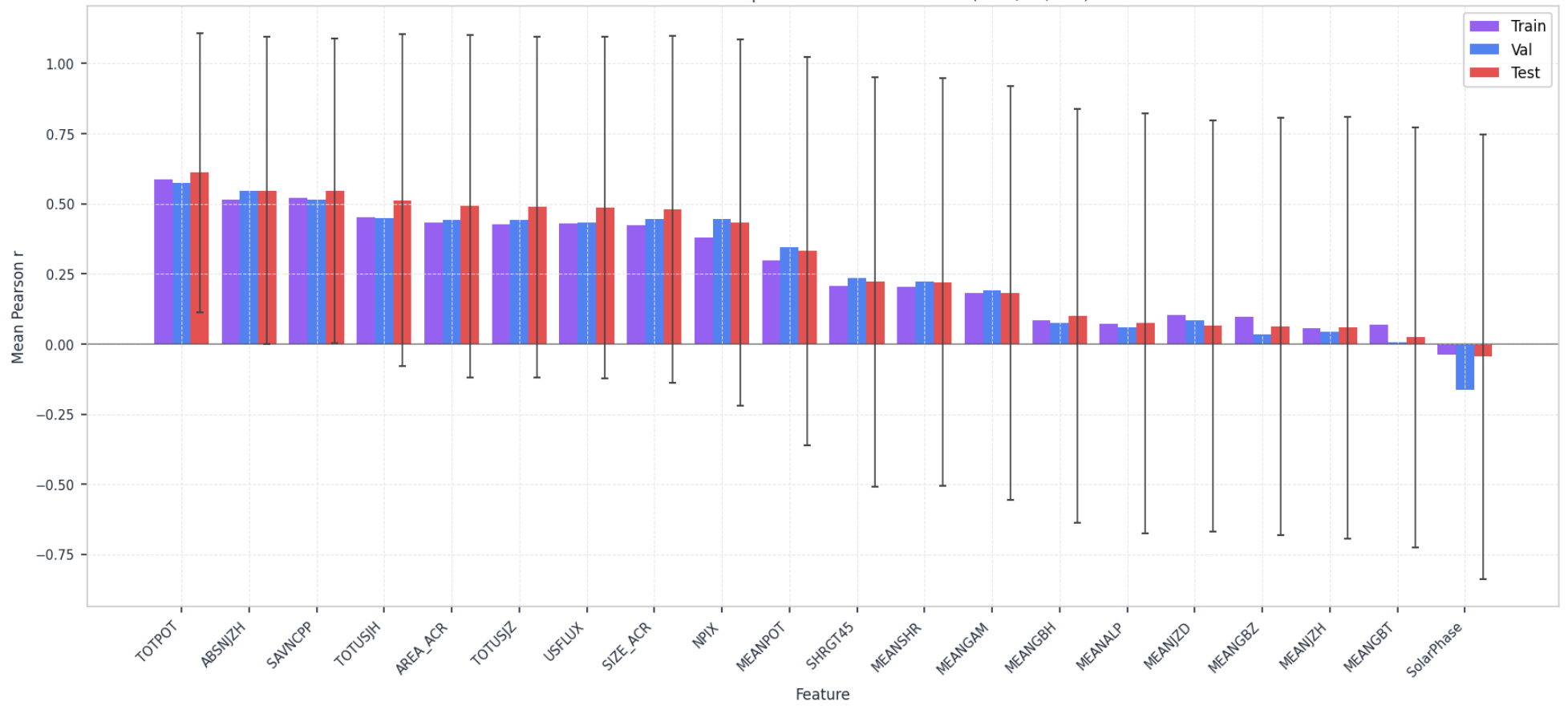


Figure 31: Overview of the strongest Pearson correlation between the second head's output and the cumulative sum of SHARP parameters for the baseline model. With the vertical axis showing the average Pearson correlation score and the horizontal axis showing each feature for the training (purple), validation (blue) and test (red) set.

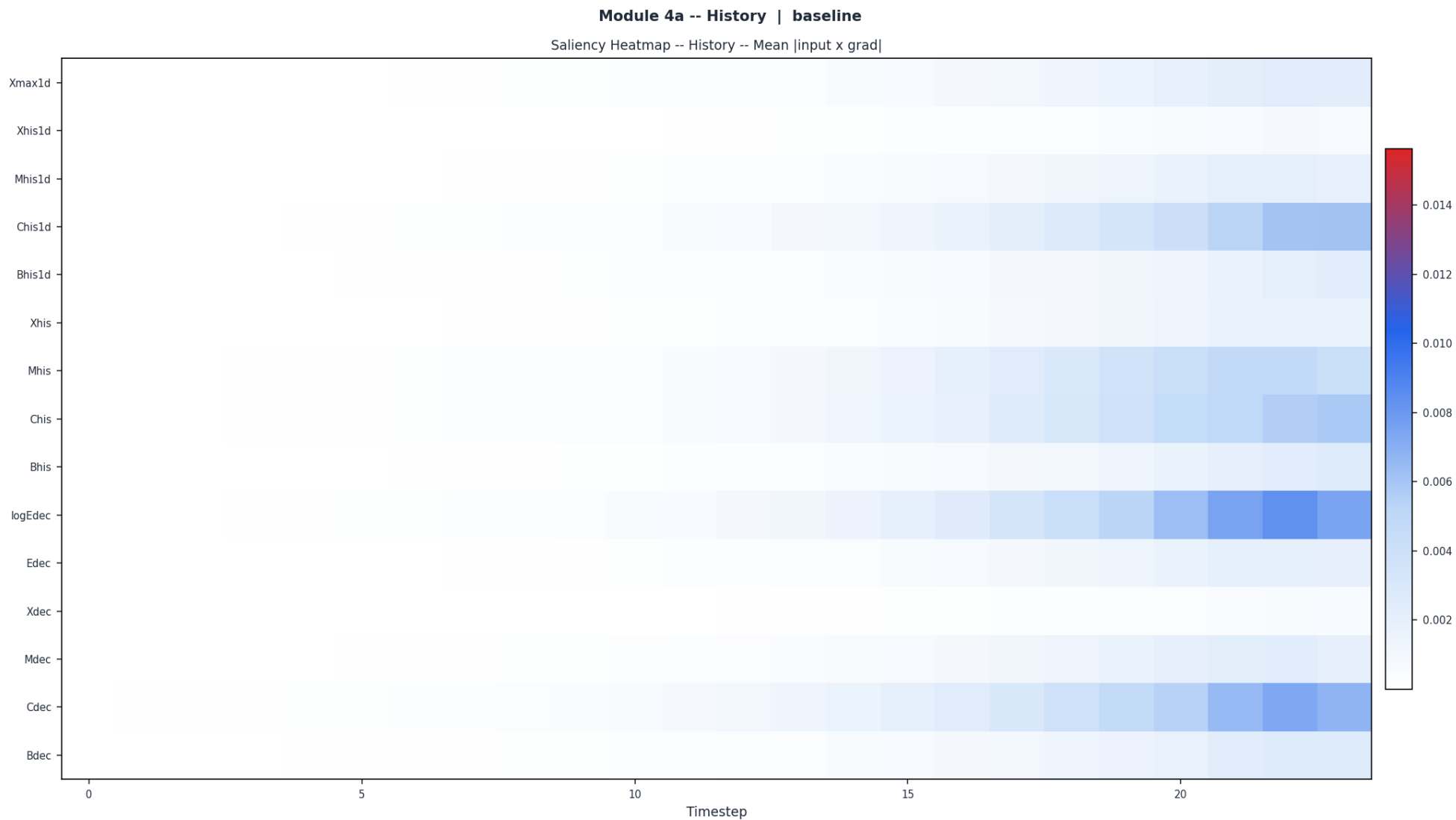


Figure 32: Saliency analysis of the baseline model for the history features. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

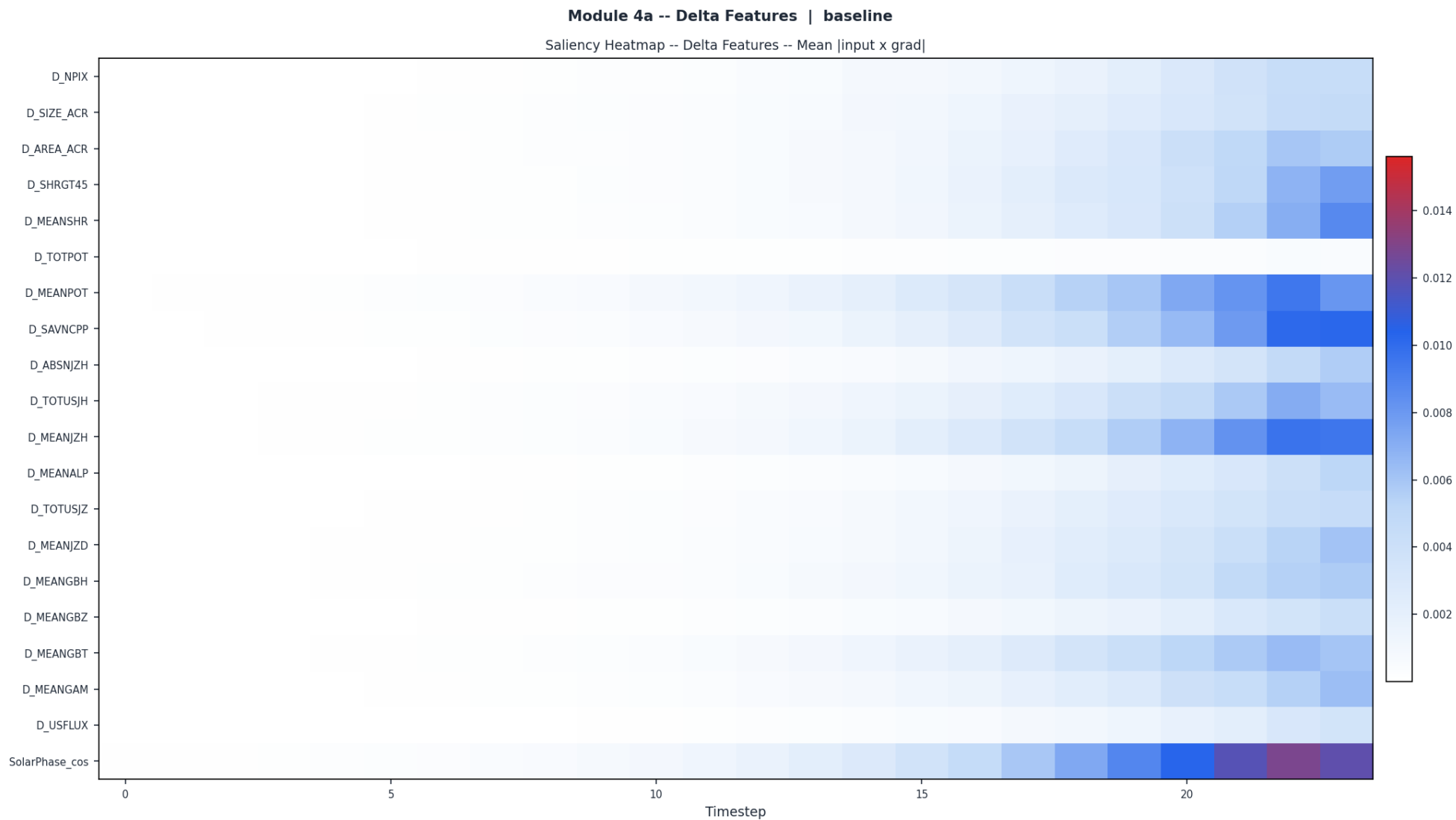


Figure 33: Saliency analysis of the baseline model for the delta features and the solar phase cosine encoding. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

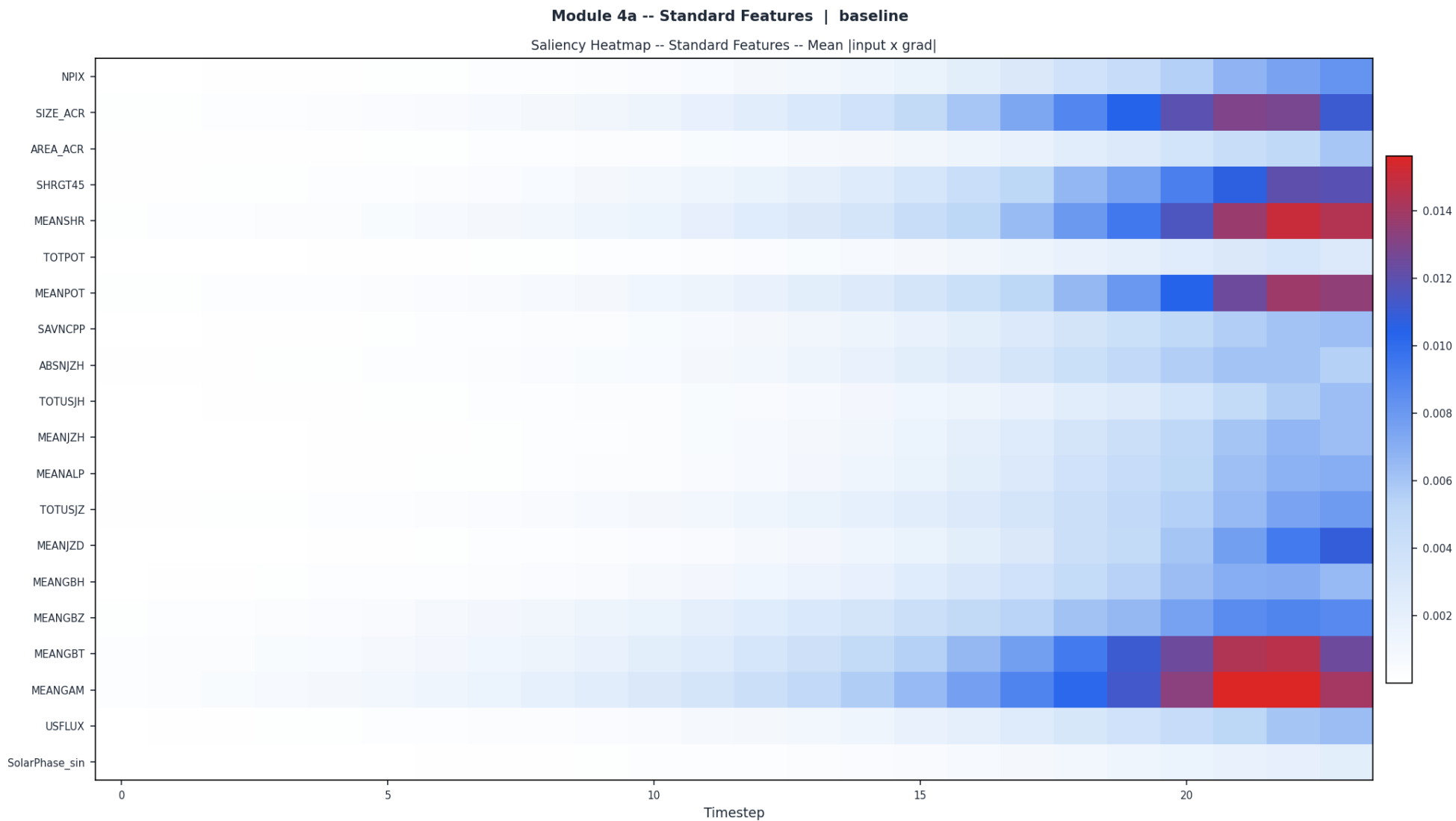


Figure 34: Saliency analysis of the baseline model for the SHARP features and the solar phase sine encoding. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Module 4a -- History | MEANGBZ_pearson

Saliency Heatmap -- History -- Mean |input x grad|

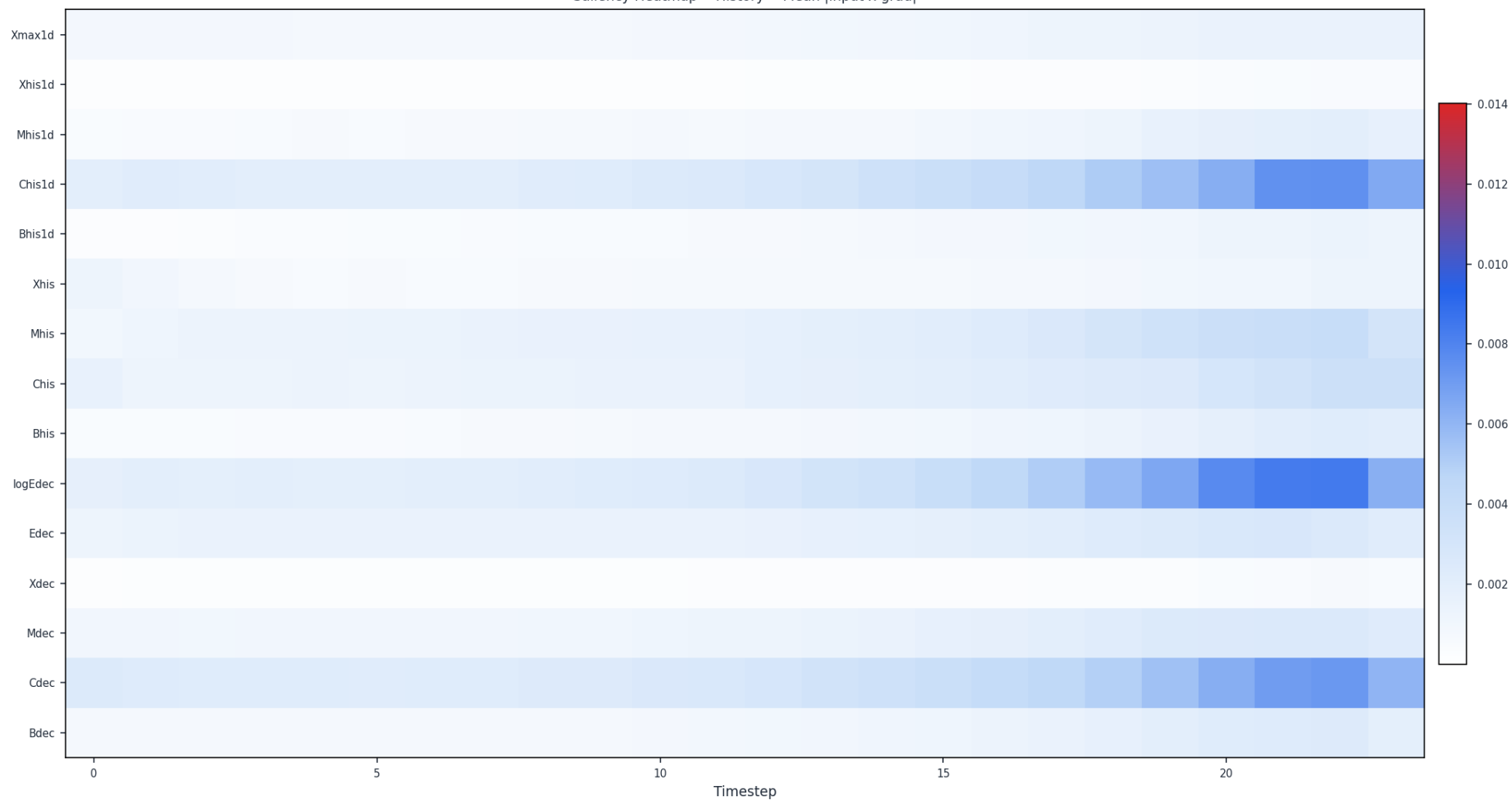


Figure 35: Saliency analysis of the Pearson MEANGBZ model for the history features . The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Module 4a -- Delta Features | MEANGBZ_pearson

Saliency Heatmap -- Delta Features -- Mean $|\text{input} \times \text{grad}|$

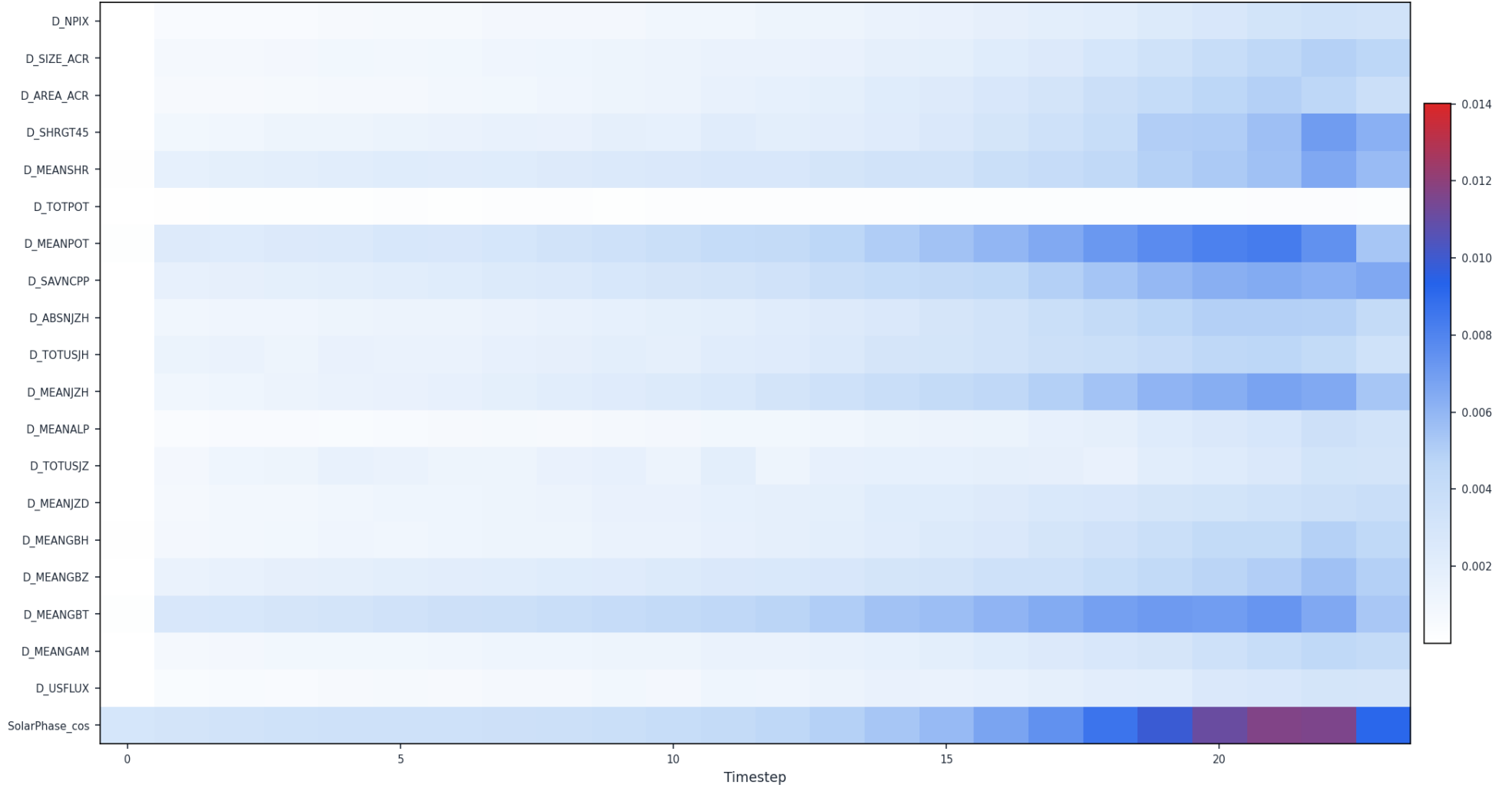


Figure 36: Saliency analysis of the Pearson MEANGBZ model for the delta features and the solar phase cosine encoding. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Module 4a -- Standard Features | MEANGBZ_pearson

Saliency Heatmap -- Standard Features -- Mean |input x grad|

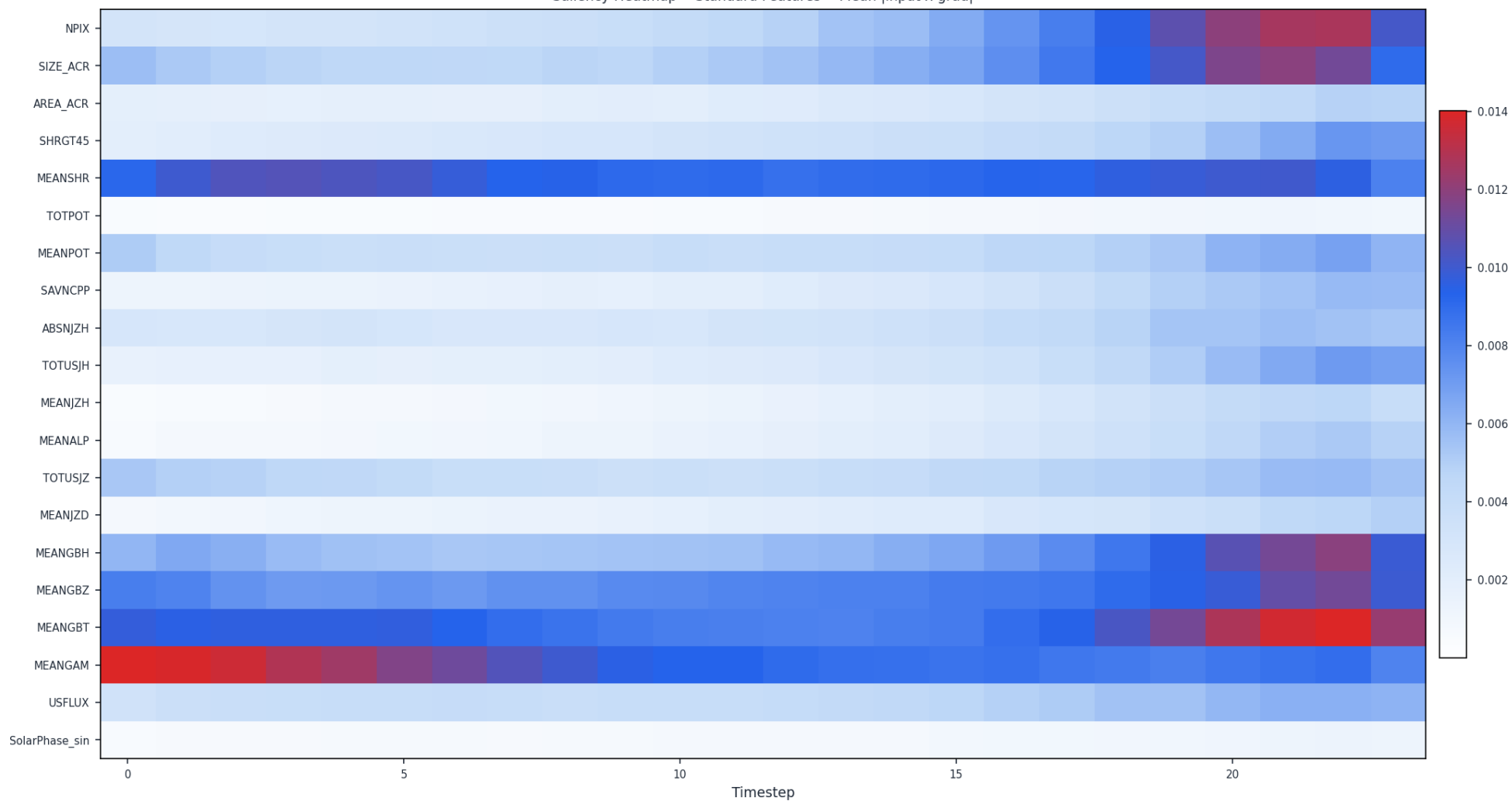


Figure 37: Saliency analysis of the Pearson MEANGBZ model for the SHARP features and the solar phase sine encoding. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Physics Correlation Analysis | MEANGBZ_pearson

SHARP Feature <-> Seq-Prob Pearson Correlation (train/val/test)

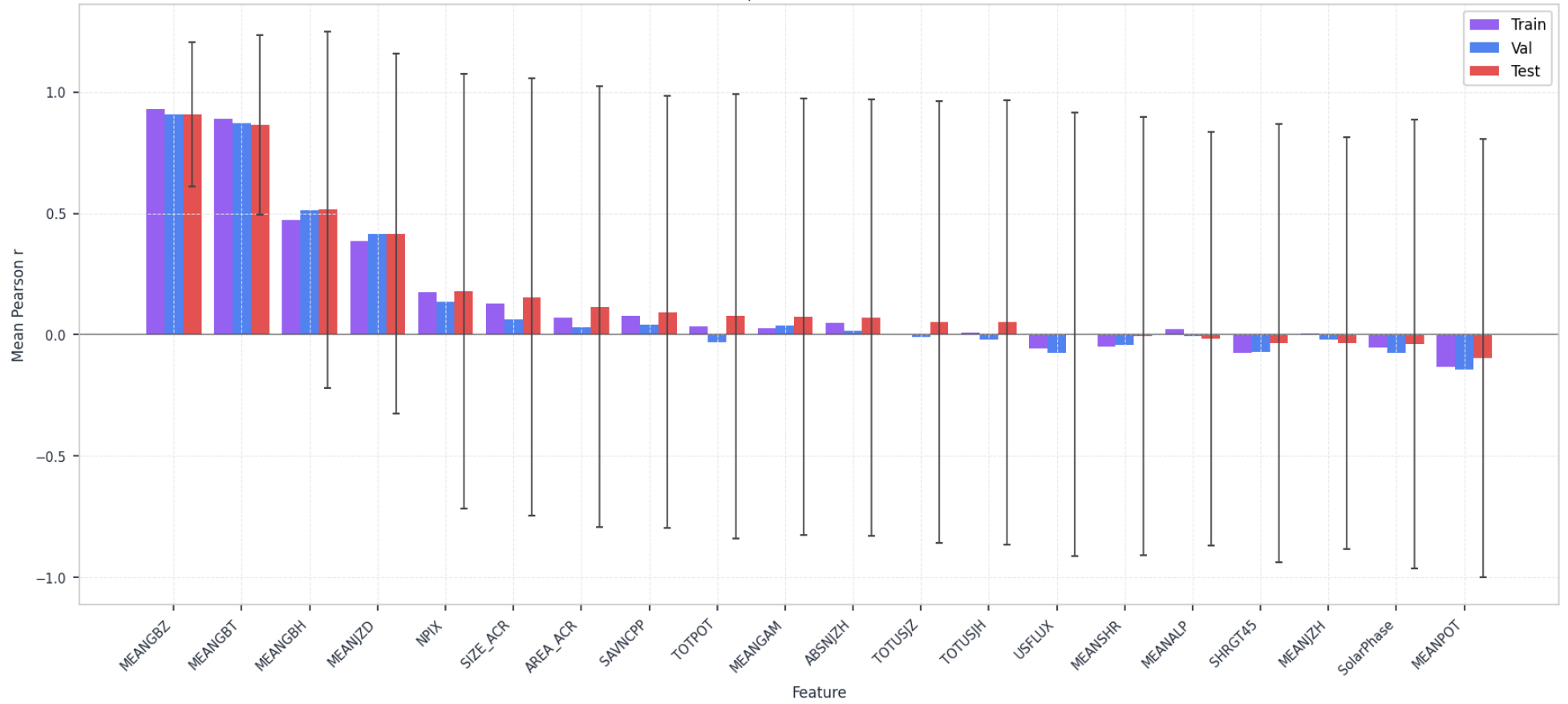


Figure 38: Overview of the strongest Pearson correlation between the second head's output and the cumulative sum of SHARP parameters for the baseline model. With the vertical axis showing the average Pearson correlation score and the horizontal axis showing each feature for the training (purple), validation (blue) and test (red) set.

Module 4a -- History | MEANGAM_kendall

Saliency Heatmap -- History -- Mean |input x grad|

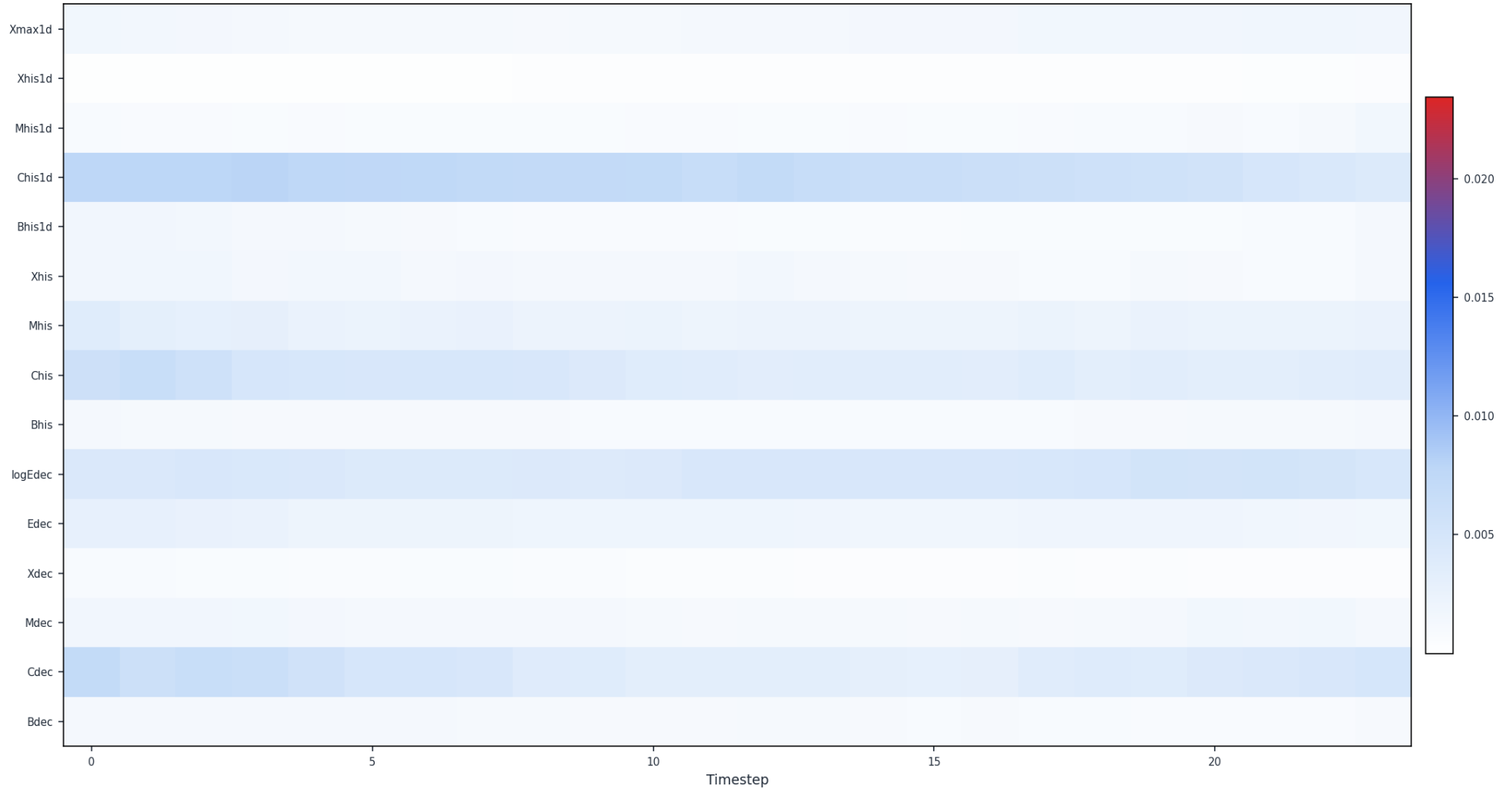


Figure 39: Saliency analysis of the Kendall MEANGAM model for the history features. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Module 4a -- Delta Features | MEANGAM_kendall

Saliency Heatmap -- Delta Features -- Mean $|\text{input} \times \text{grad}|$

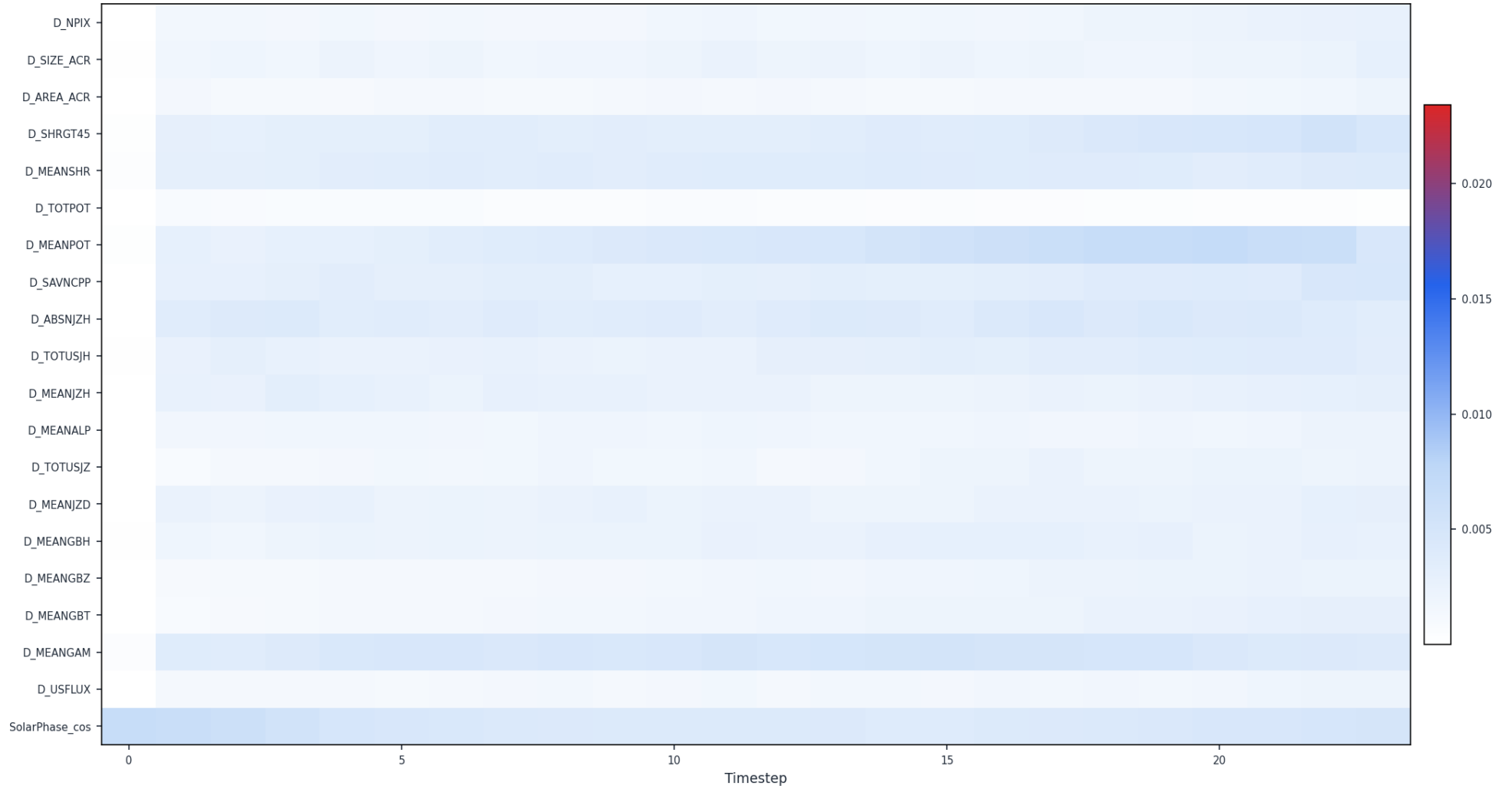


Figure 40: Saliency analysis of the Kendall MEANGAM model for the delta features and the solar phase cosine encoding. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Module 4a -- Standard Features | MEANGAM_kendall

Saliency Heatmap -- Standard Features -- Mean $|\text{input} \times \text{grad}|$

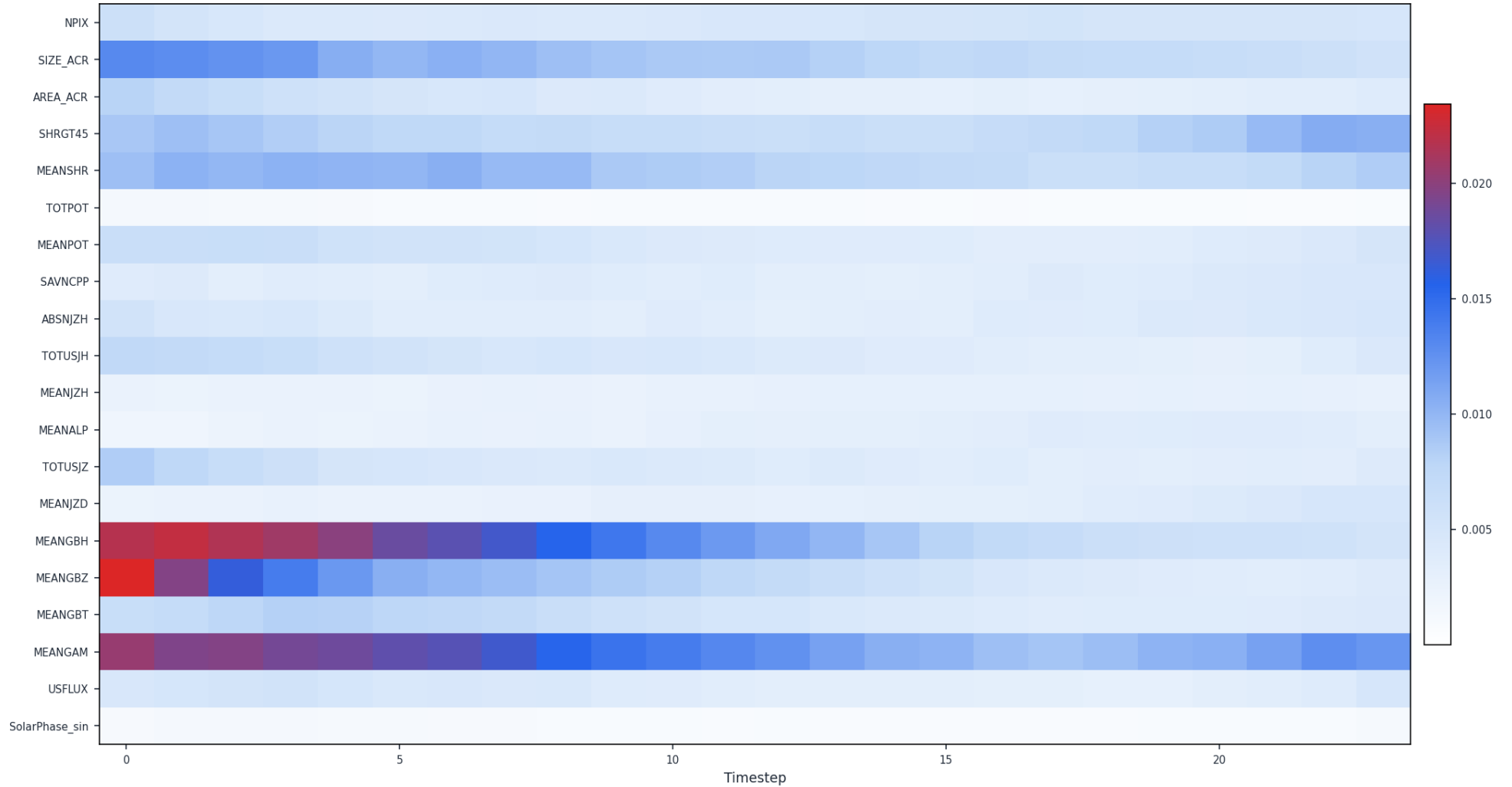


Figure 41: Saliency analysis of the Kendall MEANGAM model for the SHARP features and the solar phase sine encoding. The horizontal axis displays the hour in the sequence and the vertical axis shows the name of the feature. The color represents the value of the saliency with darker and more coloration showing stronger reliance by the model at that timestep.

Physics Correlation Analysis | MEANGAM_kendall

SHARP Feature <-> Seq-Prob Pearson Correlation (train/val/test)

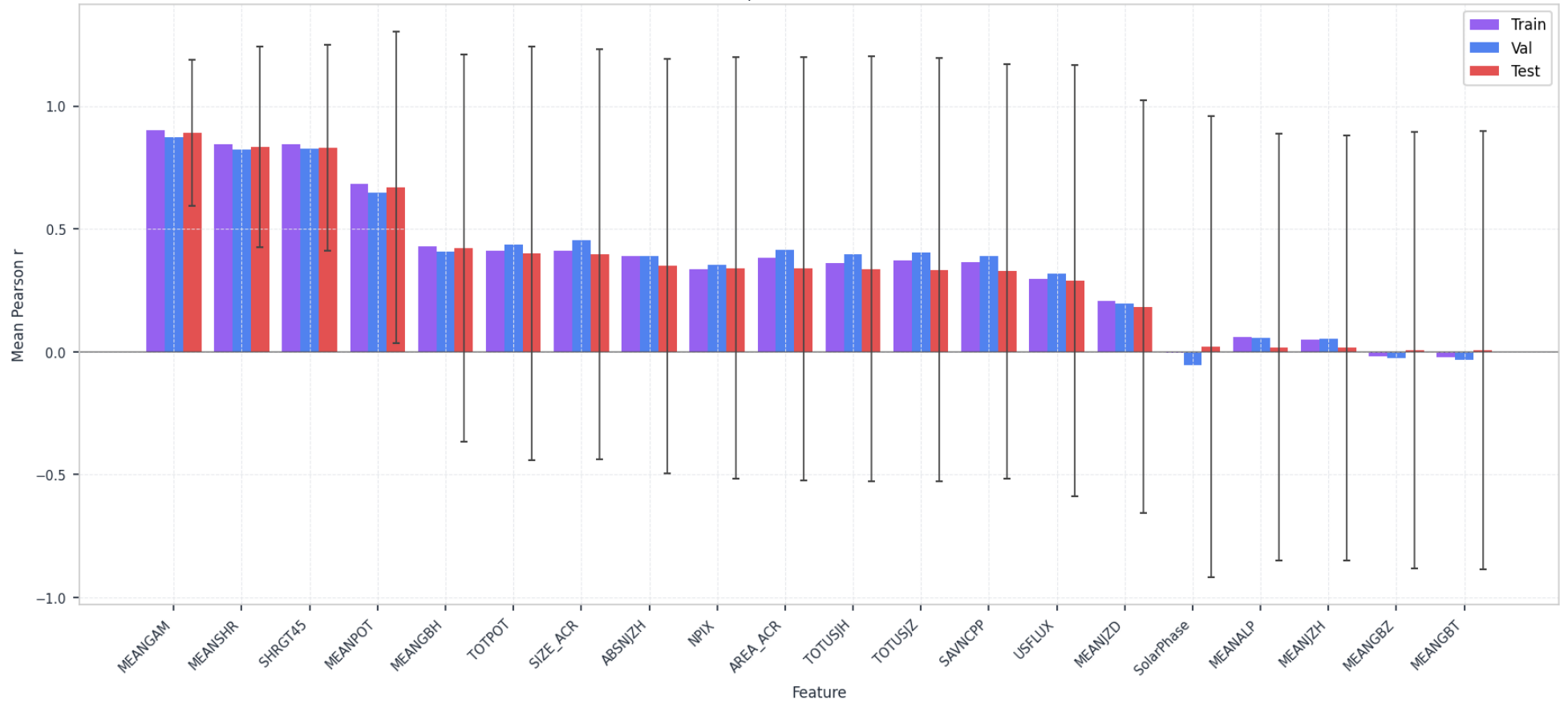


Figure 42: Overview of the strongest Pearson correlation between the second head's output and the cumulative sum of SHARP parameters for the baseline model. With the vertical axis showing the average Pearson correlation score and the horizontal axis showing each feature for the training (purple), validation (blue) and test (red) set.

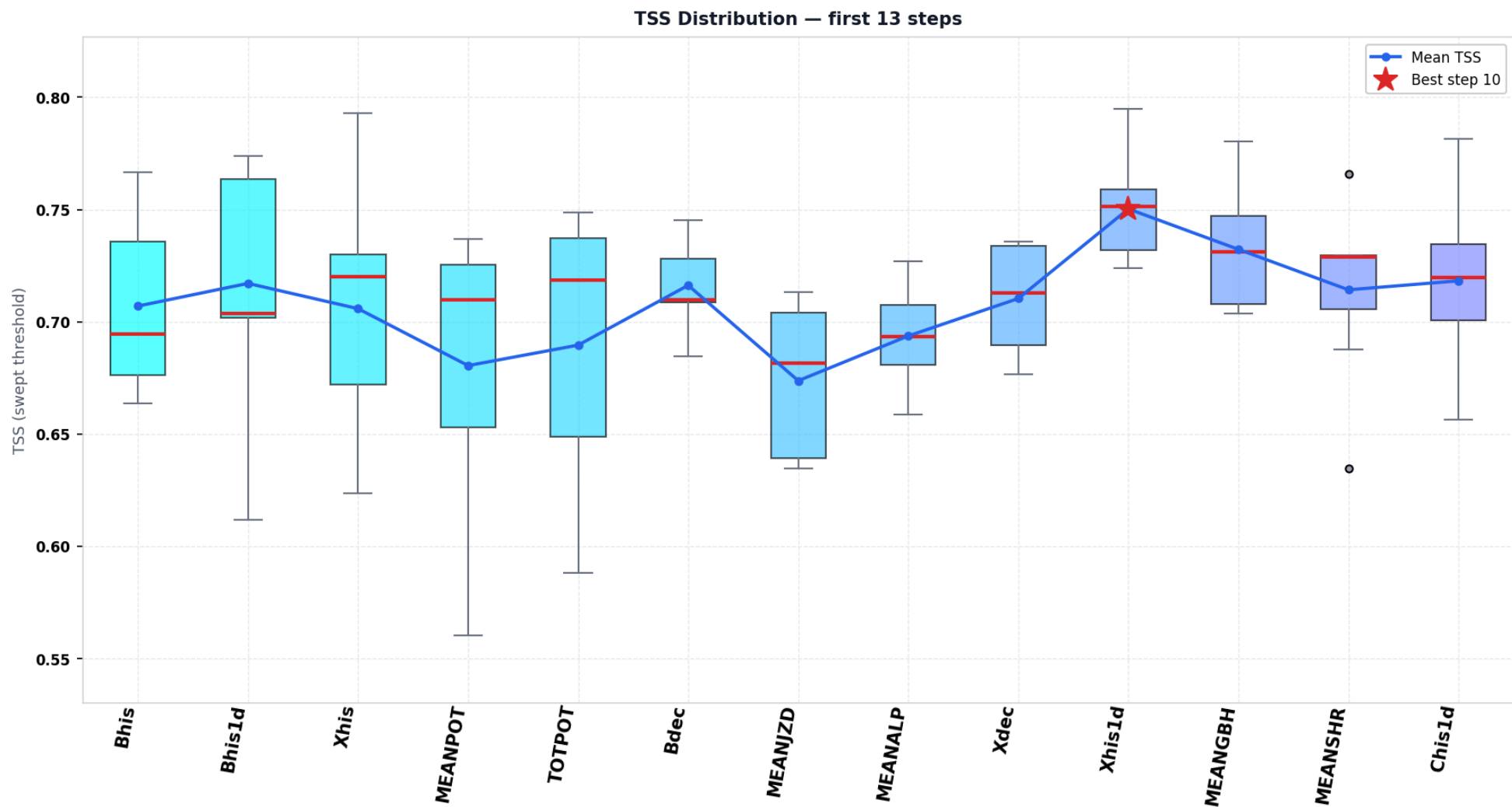


Figure 43: TSS scores for each of the first 13 elimination steps given in the form of box plots. The vertical axis shows the TSS score, where the box plots show the spread among the different data partitions used, the red line indicates the median and the blue line the average. On the horizontal axis the feature eliminated at each step is written. The step with the highest mean TSS is indicated with a star and appears at step 10 with a TSS of 0.75.

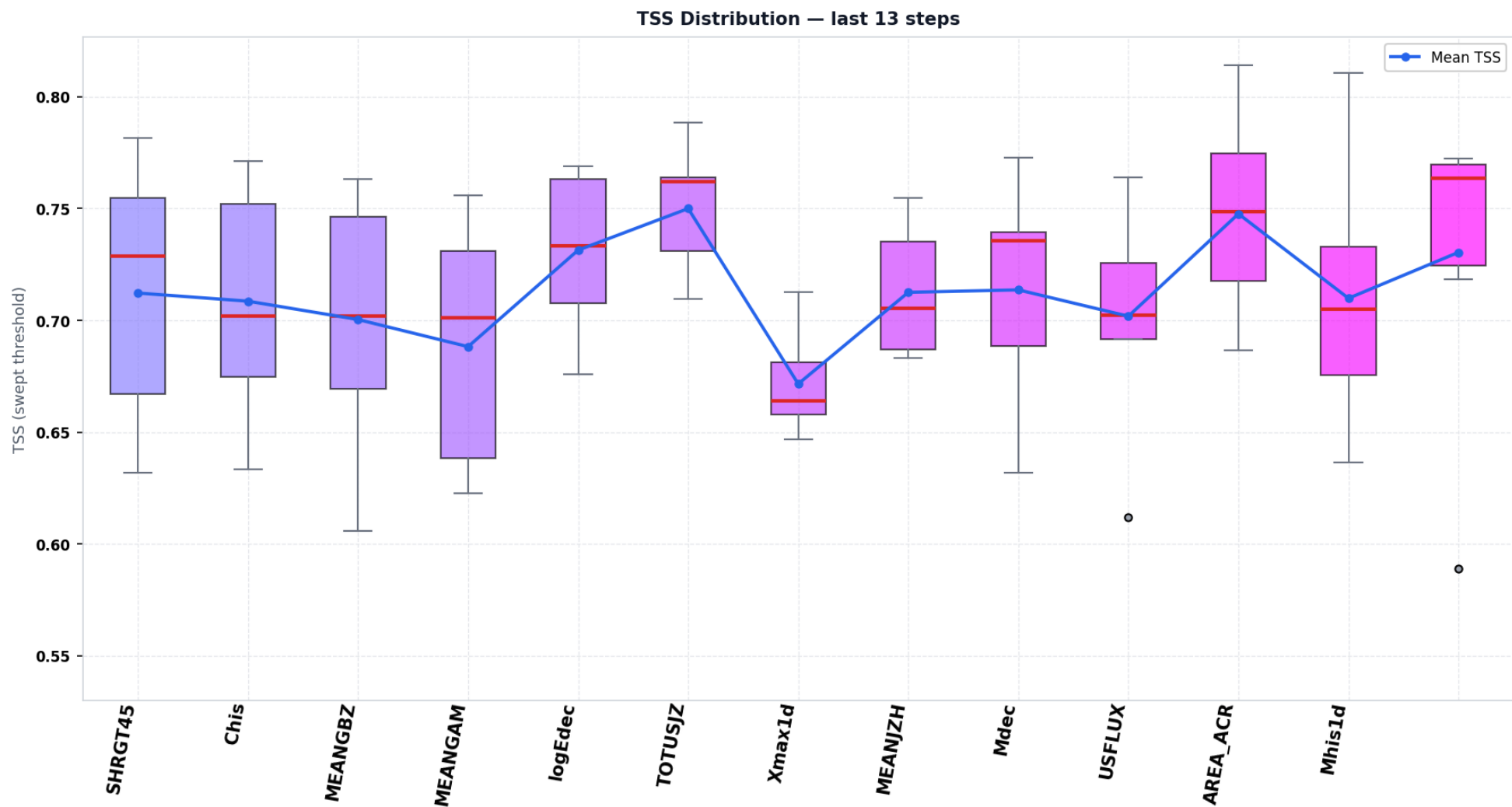


Figure 44: TSS scores for each of the last 12 elimination steps given in the form of box plots. The vertical axis shows the TSS score, where the box plots show the spread among the different data partitions used, the red line indicates the median and the blue line the average. On the horizontal axis the feature eliminated at each step is written.

Importance Score at Time of Elimination

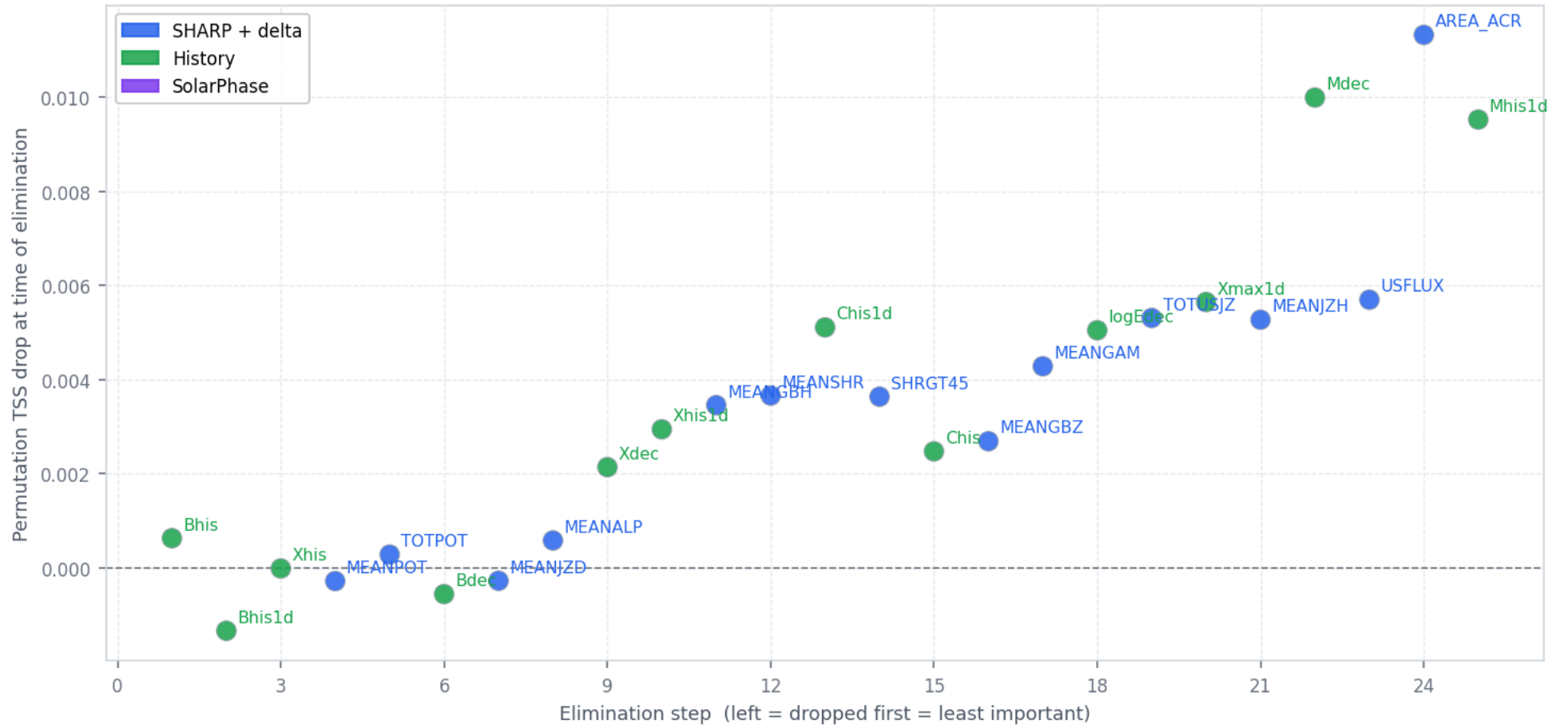


Figure 45: The vertical axis shows the change average change in TSS when permutation was done on the features depicted along the horizontal axis. Where the SHARP features are depicted blue, the history features Green and the SolarPhase survived the full elimination process and is thus not present. The dotted gray line indicates 0, meaning that any feature below this line improved the score when permutation took place.