REVIEW ARTICLE

Solar Flare Forecasting: From Data-driven Towards Physics-informed Machine Learning Models

André Leon S. Gradvohl[1,*]

[1] Faculdade de Tecnologia, Universidade Estadual de Campinas, Limeira, SP, Brazil

Abstract

Solar flares are phenomena characterized by the sudden release of accumulated magnetic energy in active regions of the solar magnetosphere. Such liberation occurs through electromagnetic radiation and high-energy particles. Flares appear as intense glows across a broad spectrum, ranging from radio waves to X- or γ-rays, and last from a few minutes to a few hours. When electromagnetic radiation reaches Earth, it can damage orbiting technologies, disrupting activities that depend on these technologies. This scoping review examines the scientific approaches to solar flare forecasting, covering methods based on physical principles, data-driven approaches using Machine Learning, and their combination in hybrid models. The text highlights the features of each approach. It argues that hybrid models, which use both observational data and knowledge of the physical nature of solar flares, offer a promising strategy. These models, known as Physics-Informed Machine Learning (PIML) models, improve accuracy, robustness, and interpretability. Key PIML strategies integrate prior physical knowledge, such as differential equations or conservation laws, by embedding them into neural network loss functions or utilizing tailored architectures. This integration supports PIML's use by enabling models that are physically plausible and less reliant on large datasets. Notably, reviewed studies show hybrid PIML models improve performance indicators, such as True Skill Statistic and False Alarm Rates, over data-driven methods, reinforcing their value for solar forecasting.

Keywords: solar flare, physics-informed machine learning, data analysis, machine learning.

1 Introduction

Discovering knowledge about nature is a complex process involving several variables that are to be observed, unfolding over time, and can include different approaches. The scientific method, established as the basis for developing the sciences in the 16th century, has four main pillars: observation, questioning, hypothesis construction, experimentation, and hypothesis analysis.

Observing a phenomenon is the basis for the discovery of new knowledge. This process includes selecting variables or attributes, collected later, and determining, from the point of view of the observer/scientist, the causes and effects of natural phenomena. When observing how a natural phenomenon occurs, the observer looks for patterns and regularities in the collected data.

Submitted: 08 October 2025
Accepted: 11 November 2025
Published: 13 December 2025

Vol. 1, **No.** 1, 2026.
10.62762/TAIS.2025.793969

***Corresponding author:**
✉ André Leon S. Gradvohl
gradvohl@unicamp.br

Citation
Gradvohl, A. L. S. (2025). Solar Flare Forecasting: From Data-driven Towards Physics-informed Machine Learning Models. *ICCK Transactions on Artificial Intelligence in Space*, 1(1), 3–24.

Subsequently, the scientist questions himself about the causes and effects of the phenomenon, builds hypotheses about its nature, and evaluates his hypotheses through experiments. In these experiments, scientists select and work on variables to establish their relevance in describing and explaining the natural phenomenon.

Scientists develop theories and models that explain natural phenomena based on observations and experiments. Therefore, scientific models, that is, simplified representations of systems, phenomena, and processes of nature, result from the process of doing Science.

Specifically in the solar flare forecasting domain, despite all the scientific research, the complex processes that culminate in a solar flare are not yet fully understood or described by Physics. The complex dynamics of the magnetic fields in the Sun's active regions, where the solar flares originate, and the exact mechanisms of storage and sudden release of magnetic energy remain topics of active investigation. This gap in knowledge makes it challenging to create models based exclusively on physical principles capable of predicting with high precision and in advance the occurrence and characteristics of solar flares.

Regarding the physical models' complexity and the growing availability of data from solar observations, machine learning techniques have emerged as a promising tool for solar flare forecasting [5]. Machine learning models can identify complex, nonlinear patterns in data, which researchers can associate with the pre-appearance of flare events, overcoming, in many cases, the performance of traditional statistical methods. Studies demonstrate the success of applying machine learning algorithms in classifying and forecasting solar flares with different intensity levels.

Nonetheless, one limitation inherent in many machine learning models, particularly in more complex architectures like deep neural networks, is their "black box" characteristic [10]. Although effective in forecasting, these models often do not clearly explain which data features are most relevant for forecasting or how the models made their decisions. This lack of interpretability makes extracting new physical knowledge from the models complex and limits confidence in their predictions in previously unobserved scenarios [42].

In this context, combining the predictive robustness of data-driven models with the interpretability and knowledge inherent in physical models is a promising path. The creation of hybrid models, which integrate both observational data and principles of physics, has the potential to improve the accuracy of solar flare forecasting and to provide valuable insight into the physical underpinnings of these phenomena. Such hybrid models can leverage the strengths of both approaches, resulting in more powerful, reliable, and scientifically enlightening prediction methods [61].

This article presents an overview of the current state of research in solar flare forecasting and points out some perspectives on injecting physics knowledge to create more robust and interpretable machine learning models. The article is organized as follows. Section 2 presents the panorama and the evolution of scientific research in solar flare forecasting using machine learning models. Section 3 generally describes models purely based on Physics, data-driven models, and physics-informed machine learning (PIML) models, referenced in this text as hybrid models. Section 4 indicates the main methods for solar flare forecasting, both those based purely on physics and exclusively on data, including a discussion on the primary assessment metrics. Section 5 maps the works of solar flare forecast using PIML and summarizes the main results. Finally, Section 6 presents the conclusions of this work.

2 The scientific interest in solar flare forecasting methods

Solar flares are sudden and intense releases of energy that occur in the atmosphere of the Sun. These solar flares release substantial amounts of radiation and energy that impact technological systems on Earth and in its near orbit. Therefore, forecasting these solar flares is important to mitigate their effects.

The study of space weather, particularly solar flare forecasting, has not only practical applications, such as protecting satellites, ensuring power grid resilience, and reducing the impacts on communication systems, but also fundamentally advances our understanding of the Sun, its internal processes, and its interaction with Earth's magnetic field and atmosphere. This deeper understanding is crucial for the ongoing advancement of basic science in this field.

As Figure 1 illustrates, since the 2000s, there has been a growing interest in developing methods for solar flare forecasting. Peaks of publications are observed between 2022 and 2024, with more than 45 publications in each of these years in both bibliographical bases, Scopus and Web of Science, considered in this

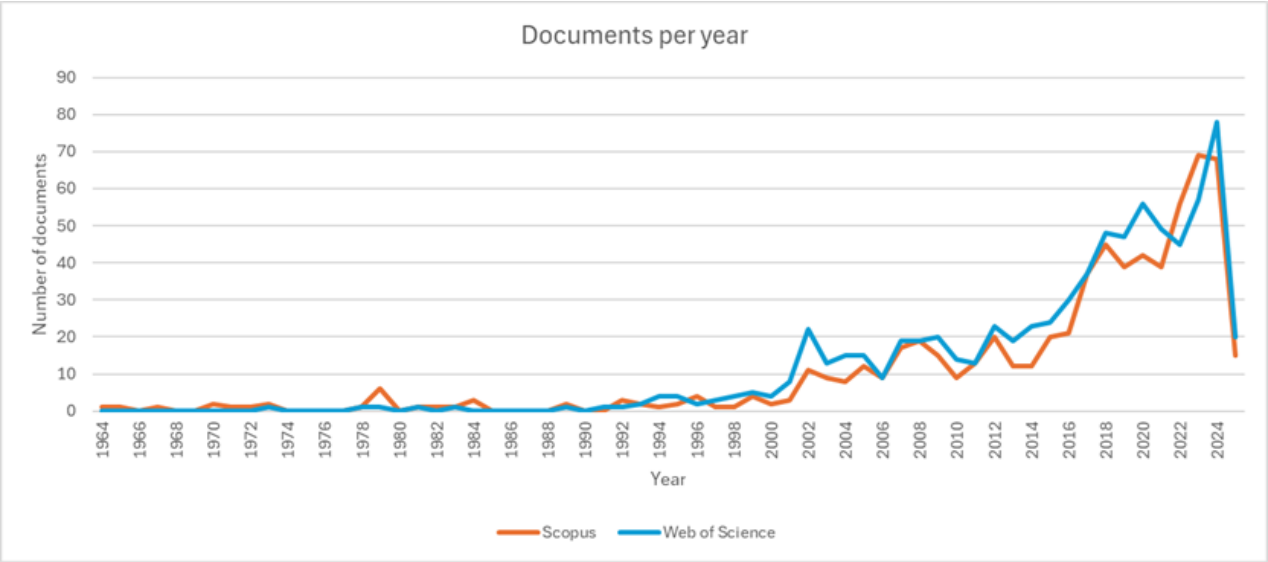

Figure 1. Number of articles published yearly, with data until March 2025. Sources: Scopus and Web of Science.

bibliographic survey. The number of publications for each year was obtained using the following query: `solar" AND flare" AND "forecasting"`.

A query designed to restrict publications to those that deal explicitly with data-driven models shows that this approach is dominant in the solar flare forecasting scenario. Figure 2 compares the number of works that explicitly mention models directed to data and works that do not mention these models. The "warmer" colors (red and orange) refer to data-driven methods, while the other colors (blue and green) refer to other approaches.

The data were obtained from 1994, as this was the year of the oldest work found [34], which involves one of the keywords related to the data-driven models. The query performed was the following: `"solar" AND "flare" AND "forecasting" AND (("machine" AND "learning") OR ("deep" AND "learning") OR ("statistical" AND "analysis") OR ("statistical" AND "methods") OR ("data" AND "driven"))`.

Figure 2 presents one interesting piece of information: the amount of works related to developing data-driven models will consistently outperform articles with other approaches from 2022 onwards, according to the Scopus and Web of Science bibliographic databases.

Figures 1 and 2 analyses were based exclusively on the Scopus and Web of Science databases. However, this analysis is representative and can be extrapolated to provide an overview of the area.

3 Models based on Physics, data-driven, and hybrid methods

In recent decades, the number of sensors and data processing equipment available has facilitated the collection and analysis of an increasing volume of data. Furthermore, the continuous development of new methods and theories in Physics, Statistics, and Computing has allowed the construction of more robust and falsifiable models, resistant to various hypothesis tests [49].

Scientific models, if well constructed, allow the prediction of the behavior of complex systems. To build these models, the greater the amount of data about the natural phenomenon, the greater the chance of having a more robust model, that is, a model that maintains stable performance and accurate predictions even when faced with uncertainties, variations, or changes in the input data.

In the current context, the creation of scientific models can be based on Physics or driven by data. The difference between physics-informed and data-driven models lies fundamentally in how they approach the scientific modeling process. Each of them will be further characterized below.

3.1 Purely physics-based models

As the name suggests, Physics-based models use knowledge of the laws of Physics, their theoretical derivations, and mathematical equations to describe the behavior of a system. Therefore, these models are built based on theoretical understanding and specialized knowledge of the domain of the

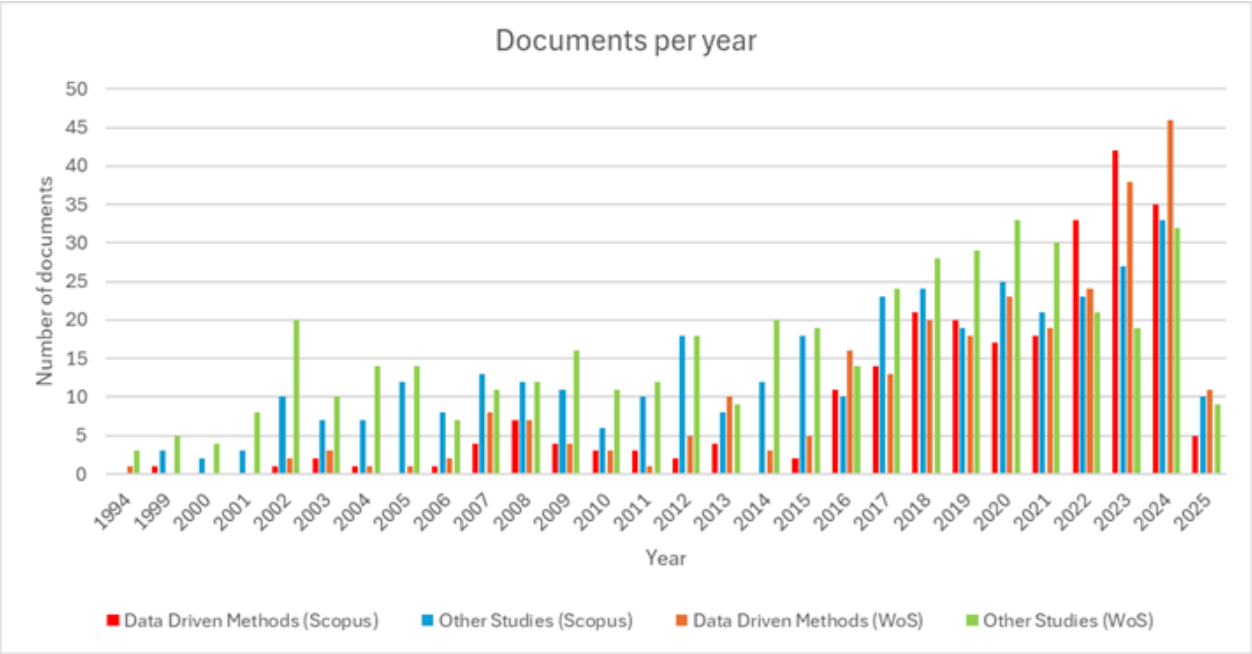

Figure 2. Comparison between the number of articles published per year dealing with data-driven and other models. Data until March 2025. Sources: Scopus and Web of Science.

phenomenon that researchers wish to model. Furthermore, scientists often express the models through mathematical equations, such as partial differential equations. After correctly modeling, the researchers compare their observations with the theoretical models, which are adjusted until they can fully explain their observations [51].

An interesting feature of physics-based models is the so-called interpretability, that is, the ability to understand and observe the internal functioning of a model and its decision-making process so that humans can understand how it arrives at a given prediction or result. In other words, an interpretable model is one whose internal logic can be traced and understood by a human analyst [6].

It is necessary to distinguish the concepts of interpretability and explainability. While interpretability focuses on the intrinsic understanding of how the model works, explainability encompasses any attempt to explain how a model makes decisions, often made a posteriori and without necessarily revealing the internal mechanisms of the model. A model's lack of interpretability and explainability characterizes it as a "black box" model [10].

However, physics-based models may have limitations. Despite their interpretability, these models may present inaccuracies or fail to consider so-called outliers, i. e., data points that lie at an abnormal distance from other values in a dataset. In other words, purely physics-based models may have difficulty dealing with noisy data. Furthermore, models of complex natural systems may only be partially known, meaning that the selected variables are insufficient to describe the natural phenomenon fully [23].

3.2 Data-driven models

In turn, data-driven models do not require extensive knowledge of the physics of the observed natural phenomenon. Those models "learn" the relations directly from the data extracted from real systems through training to find patterns in the data. The search for patterns in the data aims to find a predictive function to carry out statistical inferences or predictions. This search for data standards is called statistical learning [24].

As we can infer, the performance of data-driven models depends directly on the amount and quality of data available. Therefore, data-driven models tend to be successful and very accurate when large datasets are available for training. On the other hand, data-driven models have inferior performance with sparse data.

In contrast to physics-based models, the interpretability of data-driven models is limited. As data-driven models ignore the knowledge of domain or physical principles underlying data, it reduces interpretability, often producing unknown or inconsistent output in operational applications, despite possessing excellent functionality for

prediction, when well-trained. Therefore, data-driven models may fit the training data well, but may produce physically inconsistent predictions or be implausible when extrapolated to data not in the training set. This extrapolation to still unknown data is called generalization.

3.3 Physics-informed machine learning

In short, while physics-based models offer strong interpretability and are consistent with scientific principles, they can present limitations in forecasting accuracy and their capacity to model complex systems. In turn, data-driven models can reach high forecast accuracy since they learn directly from the data. However, they usually suffer from a lack of interpretability and strong dependence on the quality and amount of data.

To overcome the limitations of each approach, the concept of Physics-Informed Machine Learning (PIML) arose, as depicted in Figure 3. This concept integrates physics-based models with data-driven models to build physically plausible models, improving accuracy, robustness, generalizability, and interpretability, even in scenarios with scarce, noisy, or incomplete data [23, 53, 60]. Thus, including physical laws provides *a priori* knowledge that guides the learning process and reduces dependence on large datasets.

Delving deeper, the PIML incorporates the prior physical knowledge, such as partial differential equations (PDE) that describe the phenomenon, laws of energy or mass conservation, and symmetries into the machine learning process to improve the performance in tasks that involve physical mechanisms. This procedure can be implemented using various strategies, including incorporating the equations as constraints in the loss function of neural networks, designing specialized network architectures that automatically satisfy some physical boundary conditions, or using data with a physical structure to influence model convergence [23, 31, 48].

Therefore, PIML can be used to solve direct problems through the solution of PDEs, inverse problems from the inference of parameters or conditions of contour, and even to discover hidden physical relations [14].

4 Solar flare forecasting methods

Before detailing the methods for solar flare forecasting, it is necessary to establish the classes of flares. Table 1 indicates the main solar flare classes [44]. Classification occurs according to the X-ray emission peak from the solar flare, measured in a wavelength range of $0.1\,\mathrm{nm}$ to $0.8\,\mathrm{nm}$, measured in Watts per square meter ($\mathrm{W/m^2}$). Therefore, there are five classes, where classes M (moderate) and X (extreme) have the highest intensity and repercussions on Earth.

Table 1. Solar flare classes.

Class	X-ray peak flux (I) in $\mathrm{W/m^2}$
A	$I < 10^{-7}$
B	$10^{-7} \leq I < 10^{-6}$
C	$10^{-6} \leq I < 10^{-5}$
M	$10^{-5} \leq I < 10^{-4}$
X	$10^{-4} \leq I$

A solar flare occurs in an active region (AR) on the Sun with intense and complex magnetic fields, generally associated with sunspots. Therefore, active regions are groups of sunspots with a high potential to produce solar flares. The energy for solar flares is stored in twisted magnetic fields, close to or above the sunspots. That energy is released unexpectedly when these twisted magnetic fields accumulate a reasonable amount of energy [1]. Figure 4 illustrates a conceptual model of a solar flare.

According to Messerotti et al. [44], despite the efforts of the researchers, they have yet to develop equations that deterministically predict, in advance, when and where these solar flares will occur. On the other hand, several instruments in orbit around the Earth gather abundant data. Much of these are raw data, while others are images of the Sun, in different wavelengths or magnetograms.

Raw data are primary measurements obtained by scientific instruments, such as intensity and polarity of the magnetic field in different regions of the Sun, before being transformed into magnetograms, or physical parameters derivatives, or used in forecasting models [52]. A magnetogram is a graphical representation of the Sun's magnetic field variation, obtained from a magnetic dataset by a magnetometer [18].

4.1 Purely physics-based methods for solar flare forecasting

Researchers proposed several purely physics-based methods for solar flare forecasting. However, understanding the mechanisms that trigger them is still challenging [33, 38].

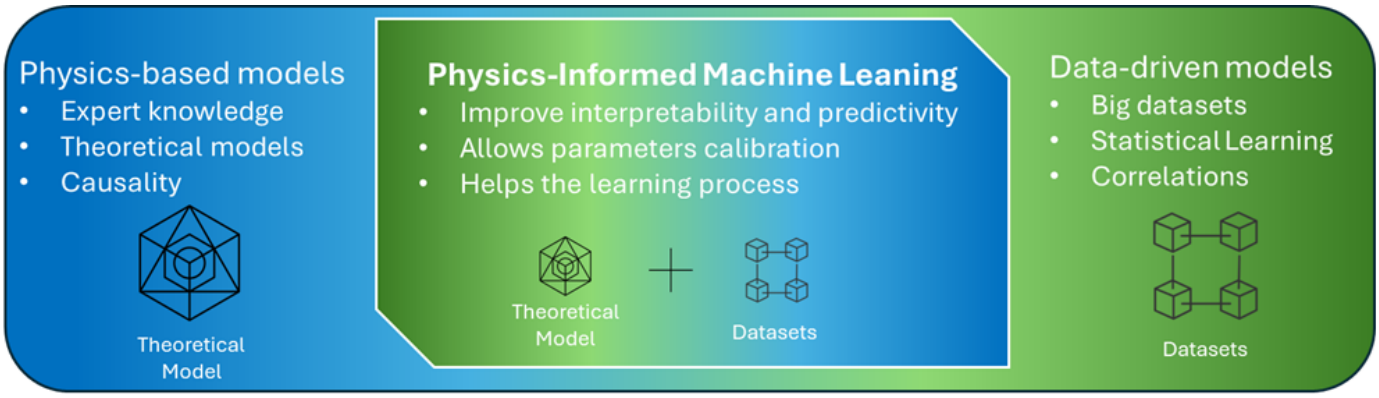

Figure 3. Illustration of the integration between physics-based models and data-driven models, resulting in the concept of Physics-Informed Machine Learning.

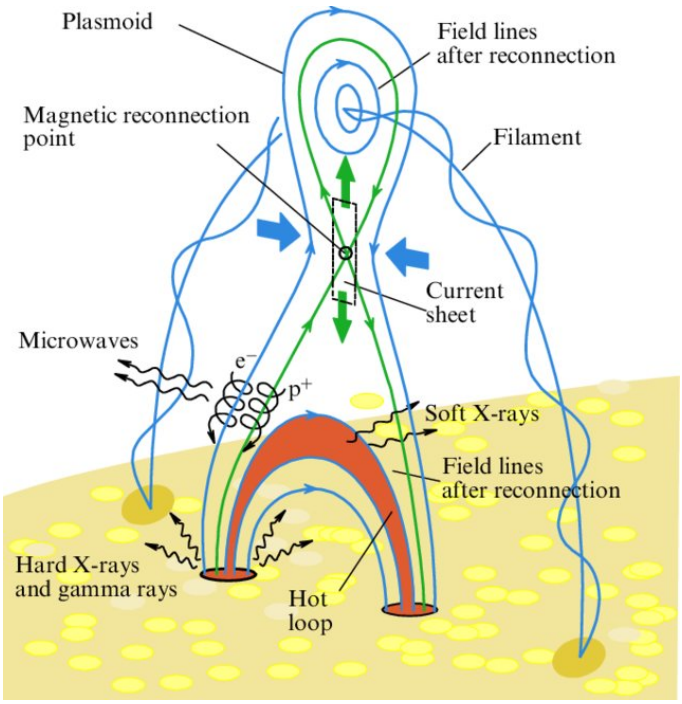

Figure 4. Illustration of the standard model of a solar flare. Source: [43].

A recent example is the k-scheme, a physics-based model proposed by [33]. This model forecasts great solar flares (class X) when a specific instability in magnetohydrodynamics (MHD) occurs, triggered by magnetic reconnection.

Thus, when analyzing X-class flares from 2008 to 2019, [33] found that the density of flow magnetic torsion, located near a line of polarity inversion on the Sun's surface, determines when and where solar flares can occur, as well as their magnitude. This method, independent of previous flare activity, can forecast imminent high-intensity flares using extrapolated nonlinear force-free magnetic field (NLFF) feature extractors.

Therefore, this method's capacity to forecast a solar flare's location and potential size beyond its immediate vicinity distinguishes it from previous empirical approaches. Although there are limitations, such as the precise determination of lead time and the difficulty in predicting flares that begin at higher altitudes, the results in this paper are promising for improving space weather forecasting.

Other methods based on magnetic topologies are also frequently used. In this case, researchers investigate the magnetic topologies of the active regions concerning the solar flares' productivity [22]. According to [44], the main proposed models are the following:

- Standard model, also called CSHKPS (name given by the initials of the authors Carmichael, Sturrock, Hirayama, Kopp, and Pneuman), which involves the Sweet-Parker/Petschek reconnection [11, 26, 32, 55].
- Flux emergence model and active region formation, which uses marginal reconnection and flux of stable current [25].
- Converging motions model, which analyzes the movements of the support points of the magnetic flux lines [50].
- Quadrupolar flare model, which analyzes converging flow patterns [27].
- Magnetic breakout model, which observes the quadrupolar magnetic configuration and considers the footpoint shearing of one side of the arcade [4].

It is important to observe that early expert systems such as Theo and Wolf, although based on expert-defined rules, incorporated knowledge on the physical

properties of the solar active regions, as the McIntosh sunspot classification and the presence of magnetic shear [44].

Despite advances, many physics-based models are still limited to subcategories of events. Hybrid forecasting models that combine features of physics-based and statistical approaches, sometimes enhanced by artificial intelligence techniques, are often used [44].

4.2 Data-driven methods for solar flare forecasting

Many data-driven methods for solar flare forecasting have been proposed and developed in recent years, as indicated in Section 2. Because they do not require very specialized knowledge of Solar Physics, as pointed out in Section 3.1, their use depends more on knowledge of Probability, Statistics, and the machine learning algorithms that will be used to find the relationships between the data. Obviously, a rich database, with an abundance of samples and balanced adequately between the solar flare classes that models want to predict, will help in the success of creating the forecasting model.

Forecasting solar flares ($\geq$ C, $\geq$ M, or $\geq$ X) using data-driven methods has focused on three main aspects, as evidenced by the research: forecast lead times or forecast horizons, flare classes or their intensities, and data handling strategies.

The most common forecast horizon is 24 hours, which provides significant preparation and response time. Some research using advanced models shows promising results with longer horizons, such as 48 hours, especially for M-class flares. Forecast classes fall into two groups: those that forecast medium- and high-intensity (M- and X-class) flares, and those that include low-intensity (A-, B-, and C-class) flares in forecasting. Most research focuses on predicting $\geq$ M, or $\geq$ X class flares, as these pose the greatest risk to terrestrial and space technologies.

Now, Section 4.2.1 provides more details on data handling strategies.

4.2.1 Data handling strategies

Key data handling strategies for solar flare forecasting with data-driven methods include data acquisition with quality control, data preprocessing and feature engineering, and addressing class imbalance.

The acquisition and quality control of data for solar forecasting models rely primarily on the Space-Weather HMI Active Region Patch (SHARP) [9], which provides high-resolution magnetogram information of Active Regions (ARs). Additionally, the Geostationary Operational Environmental Satellites (GOES) catalogs are used to record flare events.

Data quality control involves several strategies. First, longitudinal filtering excludes ARs that are too far from the solar central meridian (typically beyond $\pm 68°$) to minimize projection effects. Next, when handling missing data, one can either discard sequences with many gaps or use imputation techniques, such as hot deck or mean imputation, for small deficiencies. Finally, it is important to resolve mismatches between flare record identifiers (NOAA AR numbers) and magnetic field data (HARP patches).

A smooth transition from data acquisition and quality control leads to the next key step: data splitting and validation. These are crucial for building robust and reliable models. Typically, a dataset is split into three parts: training (70-80%), validation (10-20%), and test (10-20%) sets. The training set trains the model, the validation set tunes hyperparameters, and the test set, kept separate, provides a final, unbiased evaluation of the model's performance.

The data is typically split chronologically, with earlier data for training and later data for testing. This simulates real-world operation and prevents data leakage, where test information influences training, causing overly optimistic metrics. For a more rigorous evaluation, stratified k-fold cross-validation divides the dataset into k folds, each containing a balanced distribution of rare flare events. This provides a comprehensive assessment of the model's generalization across data subsets.

Data preprocessing and feature engineering are important strategies that make raw solar observation data suitable for forecasting models. This involves Time Series Preparation, where data is structured into sequences, such as 24 hours at 12-minute intervals. This format captures the temporal evolution of ARs. The target label is the maximum flare class in a subsequent time window, such as the next 24 hours. Magnetic parameters are then normalized or standardized using methods such as z-score scaling to ensure equal feature contributions during training.

Feature Engineering involves extracting both manual features, such as total unsigned flux and current helicity, which serve as proxies for stored magnetic energy, and automated features derived from raw magnetogram images using deep learning models.

Finally, techniques such as Dimensionality Reduction can be employed to combat overfitting, while Data Augmentation is used to enrich training set size and enhance model generalization.

It is important to highlight that solar flare datasets exhibit a significant class imbalance, where the frequent quiet (non-flaring) periods heavily outweigh the rare, powerful flare events (M and X-class), potentially biasing models toward predicting "no flare". To counter this, some strategies are employed: resampling techniques help balance the distribution by either undersampling the majority class (randomly reducing non-flaring examples) or oversampling the minority class (replicating flare examples). Alternatively, class weighting adjusts the model's loss function to impose a higher penalty for misclassifying rare flare events, forcing the model to prioritize their correct identification. Another approach is Adaptive Data Set Split, which involves creating an ensemble of classifiers by combining all positive flare samples with different partitions of negative samples.

With these data handling foundations established, the focus now shifts to methods. Data-driven methods for solar flare forecasting can be categorized into two groups: classification and regression. The following sections address each group.

4.2.2 *Classification task methods*

For solar flare forecasting, the classification tasks aim to categorize future solar flares' occurrence or features into distinct classes. Thus, the classification search to respond to questions as "one solar flare of a certain class (e.g., M or X) will occur within a certain period (e. g., 24, 48, or 72 hours)?" This type of question implies a binary classification, i. e., the occurrence of a solar flare in the class target (classes M or X) versus its non-occurrence. Therefore, the objective is to predict whether one solar flare of a certain magnitude will occur.

Researchers employ a variety of machine learning techniques to perform these classification tasks. Some examples include:

- K-Nearest Neighbors (k-NN) [15].
- Support Vector Machine (SVM) algorithms [8, 17, 46, 52], or Relevance Vector Machine (RVM) [58].
- Learning Vector Quantization (LVQ) [37].
- Decision tree-based algorithms, such as Random Forest [39], Light Gradient Boosting Machine (LightGBM) [52], or Extreme Gradient Boosting (XGBoost).
- Neural Networks, such as Multi-layer Perceptrons [15] and [17]; and Deep neural networks [57], such as Long-Short Term Memory (LSTM) [40], Long-Term Recurrent Convolutional Neural Networks [21], deep convolutional neural networks (CNN) [18], and Patch-distributed CNN [16].

In some works, using an ensemble of methods helps improve predictions. In machine learning models, the ensembles are a set of classifiers or models whose outputs are combined to generate a more robust classifier or forecast [52]. The main objective of the ensembles is to optimize the predictive performance, often minimizing the noise of models. Applications of ensemble methods include random forests and Gradient Boosted Tree Models (such as LightGBM and XGBoost).

The ensembles can use various strategies to combine predictions. Some combinations include the following [19]:

- Voting majority (hard voting): the final output is determined for the class or value predicted by the most of individual models.
- Average probabilities (soft voting): The probabilities predicted by each model are averaged to obtain a final probability.
- Linear combination: The predictions from each member are multiplied by specific weights and summed to produce the ensemble prediction.

Weights can be determined by several approaches, such minimizing error variance and optimizing metrics with or without constraints. An important advantage of ensemble models is the ability to estimate uncertainties from the forecast, even when none of the individual members have uncertainties associated with them. The main sources of uncertainty in multi-model ensembles are statistical (quantified by the weighted standard deviation) and systematic (depending especially on the uncertainty of the combination weights).

4.2.3 *Metrics for performance assessment of classification models*

The performance evaluation of these classification models uses metrics specific to classification problems, especially in scenarios where the classes are imbalanced. For example, high intensity solar

flare classes (M and X) are less frequent than classes A, B, and C (low intensity).

The parameters to calculate these metrics consider the number of predictions correctly classified as positive (True Positive – TP), correctly classified as negative (True Negative – TN), incorrectly classified as positive (False Positive – FP), and incorrectly classified as negative (False Negative – FN). From these parameters, the following fundamental metrics can be calculated:

- Accuracy (ACC) is the percentage of hits in relation to the total number of predictions. Equation 1 describes the accuracy.

$$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{1}$$

- Precision (PRE) indicates the proportion of instances correctly expected as positive in relation to the total number of instances classified as positive. Equation 2 calculates the precision.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}. \tag{2}$$

- False Alarm Ratio (FAR) informs the fraction of predictions associated with the non-occurrence of the events. Equation 3 describes this metric. Note that the FAR metric is complementary to the precision metric.

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}} = 1-\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\right) = 1-\mathrm{PRE}\,. \tag{3}$$

- True Positive Rate (TPR), recall, or sensitivity, evaluates the method's ability to detect results classified as positive successfully. This metric is described in Equation 4.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}. \tag{4}$$

- True Negative Rate (TNR), or specificity, evaluates the capacity to detect negative results. Equation 5 calculates this metric.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{(\mathrm{FP}+\mathrm{TN})}. \tag{5}$$

Datasets with solar flares are naturally imbalanced. High-intensity solar flares (classes M and X) are rare and are precisely the ones that need to be predicted. Therefore, in addition to the previous metrics, other metrics that better address imbalanced databases can be used. These metrics are as follows:

- F1 Score is a harmonic mean calculated based on precision (PRE) and sensitivity (TPR). Thus, this metric considers false negatives (FN) and false positives (FP), and not only the number of incorrect predictions. Therefore, unlike accuracy, which is sensitive to class distribution, the F1 Score provides a balanced measure that considers both false positives and false negatives, making it more suitable for imbalanced datasets. Equation 6 describes how to calculate the F1 Score.

$$\mathrm{F1} = 2\times\frac{\mathrm{PRE}\times\mathrm{TPR}}{\mathrm{PRE}+\mathrm{TPR}} = \frac{2\times\mathrm{TP}}{2\times\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}. \tag{6}$$

- True Skill Statistic (TSS) is a metric that accounts for true positive rates (TPR) and true negative rates (TNR). In other words, TSS computes the difference between the probability of detection (TPR) and the probability of false detection (TNR) [29]. It prioritizes both positive and negative successes, making it widely used for evaluating solar flare forecasting methods. Its adoption in solar flare forecasting is due to its insensitivity to class imbalance, as the M and X classes are inherently rare. TSS is also valued for being unbiased toward event or non-event ratios. Bloomfield et al. [7] proposed and rigorously justified its use. Later, Leka et al. [35] reaffirmed this rationale in operational benchmarking studies. However, Doswell et al. [13] note that TSS can be vulnerable for rare event prediction, as it tends to match the TPR when correct non-occurrence forecasts dominate, which often happens with high-intensity solar flares. Equation 7 describes the TSS metric.

$$\begin{aligned}\mathrm{TSS} &= \mathrm{TPR}+\mathrm{TNR}-1\\ &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}.\end{aligned} \tag{7}$$

- The Heidke Skill Score (HSS) metric has gained greater relevance in recent work, as it indicates how good a forecast is in relation to a random prediction [29]. Values range from $-\infty$ to 1, and the closer to 1, the closer the model is to perfect performance. A close value of zero indicates that

the model is as good as a random prediction. Doswell et al. [13] recommend the use of HSS for predicting rare events. Equation 8 describes the calculation of HSS.

$$\mathrm{HSS} = \frac{2\times(\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN})}{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{FN}+\mathrm{TN})+(\mathrm{TP}+\mathrm{FP})\times(\mathrm{FP}+\mathrm{TN})}. \tag{8}$$

- The area under the receiver operating characteristic (ROC) curve (AUC) measures how well a model can distinguish between positive and negative samples. It is a key metric to assess the performance of diagnostic tests and binary classifiers. The ROC curve shows the performance of one binary classifier with different decision thresholds. It plots the rate of true Positive (TPR) against the rate of False Positive ($\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$) in function of the decision thresholds, as Figure 5 illustrates. Thus, when AUC is closer to 1, the model's performance is better. However, the closer the model is to 0.5, the closer it is to a random prediction.

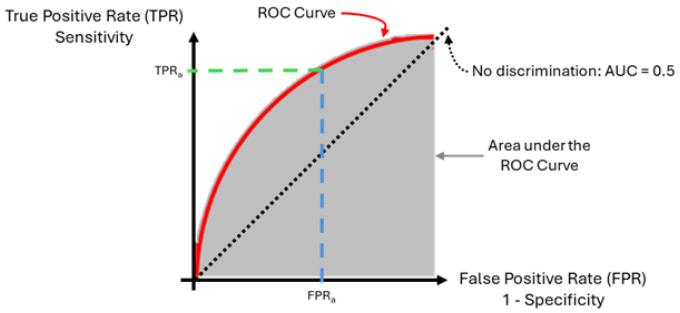

Figure 5. ROC curve.

Despite the number of metrics available, presenting results in more than one metric is always important. Some metrics evaluate better forecasting methods with imbalanced data, and others that prioritize hit or error rates. Thus, presenting the results using several metrics will allow for a more robust analysis.

4.2.4 *Methods for regression tasks*

Regression is a supervised machine learning method used to predict continuous values. The ultimate goal of a regression algorithm is to plot a best-fit line or curve across the data. This line or curve will show the trend and the predicted values.

In solar flare forecasting, the regression task aims to predict a single numeric continuous value representing a feature of a future solar flare, such as the maximum intensity (for example, the peak of the X-ray flux) or its duration. Unlike classification tasks, which categorize the occurrence or characteristics of flares into distinct classes, regression seeks to estimate the magnitude of a future event.

Although much of the work concentrates on solar flare classification, the regression tasks aim to answer questions like:

- What will be the peak of X-ray flux (in $\mathrm{W/m^2}$) for the next solar flare?
- How much time (in minutes or hours) will the next solar flare last?
- What will be the area of the active regions that can produce a solar flare? In the end, although the relationship between the area and the magnitude of the solar flare is not linear, the area can indicate the solar flare's intensity.

The main regression models are the following:

- Linear regression: establishes a straight line of tendencies to forecast values, as shown in Figure 6.
- Polynomial regression: plots a curve between the points (data), as shown in Figure 7.
- Logistic regression: determines if points (data) are below or above a line. Note, however, that this type of regression is used in the classification task.

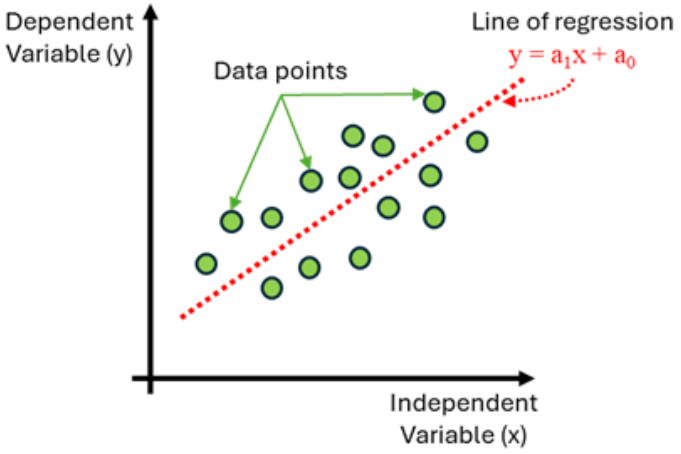

Figure 6. Linear regression.

4.2.5 *Metrics for regression models assessment*

In this section, we present the primary metrics used in evaluating the performance of regression models. Most of the metrics evaluate the error, that is, the difference between each one of the target values predicted (P_i) and the observed values (O_i) within a data set with n samples n ($i = 1, \ldots, n$).

The Mean Absolute Error (MAE), described by Equation 9, calculates the average absolute difference

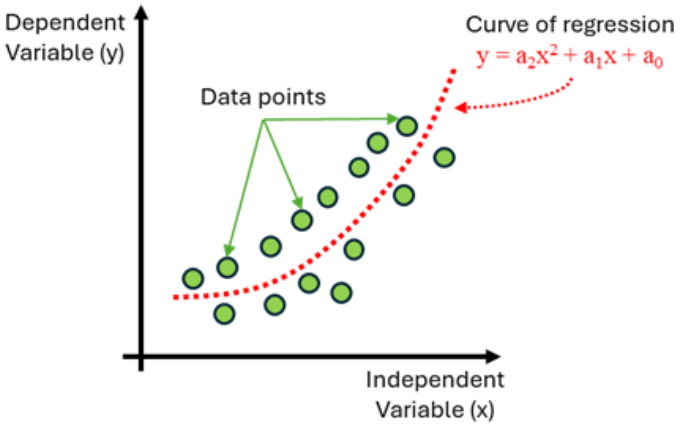

Figure 7. Polynomial regression.

between the predicted target values (P_i) and the observed value (O_i).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|O_i - P_i| \tag{9}$$

In turn, the Mean Square Error (MSE) evaluates the average squared difference between the predicted and observed values in the dataset [30]. Equation 10 describes the calculation of MSE, which shows how close a regression line is to a set of data points. Furthermore, MSE penalizes larger errors more, since the equation involves the elevation to the power of two of the differences. Therefore, the bigger the value calculated, the worse the model's performance.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2 \tag{10}$$

As the MSE, the Root Mean Square Error (RMSE) also calculates the difference between the target value predicted (P_i) and the observed value (O_i). The difference between the two metrics is in applying the square root, as Equation 11 shows. Applying the square root keeps the result on the same scale as the original data and allows a better interpretation of the error's result.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} = \sqrt{\text{MSE}} \tag{11}$$

Variance and bias metrics evaluate regression models. Consequently, there are two types of undesirable situations: overfitting and underfitting, which we will discuss next.

A bias occurs when a model's predictions are systematically wrong due to incorrect assumptions. Bias can also be defined as the error between the model's average prediction and the actual observed (true) data. Therefore, the lower the bias, the better. Equation 12 illustrates the concept mathematically. Consider that Y is the true value of a parameter, $\hat{Y}$ is one estimator of Y based on a sample of data, and $E(\hat{Y})$ is the expected value of the estimator $\hat{Y}$.

$$\text{Bias}(\hat{Y}) = E(\hat{Y}) - Y \tag{12}$$

On the other hand, the variance indicates how the model's predictions vary due to using different subsets of training data. In other words, it measures how much a model's predictions change when trained on different data. High variance means the model fits the training data nicely but does not generalize well to new data. Therefore, the lower the variance, the better. Equation 13 illustrates the concept mathematically. In this case, also consider that $E[\hat{Y}]$ is the expected mean of the predicted values.

$$\text{Variance} = E[(\hat{Y} - E[\hat{Y}])^2] \tag{13}$$

One essential concept to note is the tradeoff between bias and variance. If the generated model is too simple, it can have high bias and low variance. On the other hand, if the model is very complex, it may have high variance and low bias. Figure 8 shows the relationship between these two metrics. It is crucial for the model's success to seek a balance (tradeoff) between bias and variance.

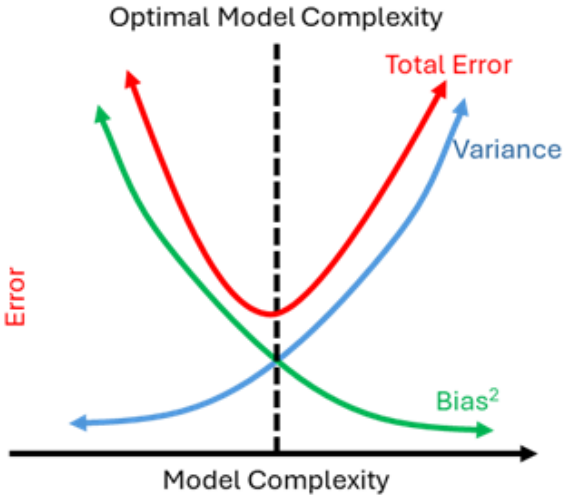

Figure 8. Relationship between bias and variance considering the complexity of a model.

Regarding undesirable situations, overfitting occurs when the machine learning model provides fully accurate predictions on the training data, but these predictions fail on new data. In this case, there is high variance and low bias. Therefore, the model cannot generalize to other data. The leading causes of overfitting are:

- A small set of training data without a sizable number of samples to represent all the possible data.
- A dataset for training is very noisy, with lots of irrelevant information.
- A high complexity model that learns with noisy data.

To address overfitting, the literature offers strategies that are presented next.

- Regularization Techniques:
  - Dropout randomly disables some neurons during training to prevent over-reliance and encourage robust learning [59].
  - L1/L2 Regularization adds penalties to cost functions to discourage large or unnecessary weights, leading to simpler models and reduced memorization. Specifically, L2 (Ridge) regularization discourages large weights by shrinking them towards zero, while L1 (Lasso) regularization can set some weights exactly to zero. Both approaches result in a simpler model that is less likely to memorize data [45].
- Data Augmentation creates more samples of the minority class, using techniques such as Synthetic Minority Over-sampling Technique (SMOTE) [12]; or image transformations [18], to reduce model bias.
- Training Strategies:
  - Early stopping halts training when validation error increases, saving the model at its best generalization point [54].
  - Cross-validation, particularly important in time series analysis, tests the model by sequentially validating it on future periods. It trains the model to generalize better when making predictions on unseen data [47].

Underfitting occurs when the model is so simple that it cannot fit the relationships between the data (the input variables and the target values) even during the training phase. In this case, there is high bias and low variance. The leading causes of underfitting are:

- Insufficient or very noisy data.
- Inappropriate preprocessing.
- Insufficient time for training.
- Simplicity of the model.

To mitigate underfitting, it is necessary to increase the model's capacity to learn complex, non-linear patterns. The literature supports the following strategies [3].

- Enhance model complexity by utilizing sophisticated models, including non-linear approaches and deep learning architectures. CNNs can extract spatial features from solar images, and LSTMs can model temporal evolution, effectively capturing data complexity.
- Apply feature engineering by incorporating parameters like magnetic gradients and non-neutrality measures. In particular, using physical knowledge simplifies the problem and enables the model to learn patterns more effectively.
- Reduce regularization by lowering L1/L2 penalties or decreasing dropout rate to avoid over-simplification.
- Train for longer periods or optimize with better-tuned parameters to give the model more time to learn effectively.

5 Mapping of solar flare forecasting work using PIML

As reported in Section 2, research has used machine learning and deep learning algorithms to create models for predicting solar flares. Excelent work has achieved high accuracy, above 85%, with positive hit rates. Despite the superior results, some doubts arise. Among them are:

- Could we improve solar flare forecasting driven solely by data if we add information about the physics of these phenomena?
- Are there works using the PIML for solar flare forecasting (PIML-SFF)?
- What results did these works achieve using PIML-SFF?

- What challenges do researchers using PIML-SFF face?

Considering those questions, an exploratory study was conducted on using PIML for solar flare forecasting. Section 5.1 presents the details of the literature survey for this exploratory study.

5.1 Parameters for the bibliographic survey and quantitative results

Before carrying out the bibliographic survey, preliminary tests were conducted considering keywords relevant to the search in the main bibliographical databases. The following keywords were selected: "`physics informed`", "`physics based`", "`solar flare`", "`forecasting`", "`prediction`", "`machine learning`".

To the surprise, the number of publications obtained when using some combinations of these keywords was minimal compared to expectations. In the initial queries, without restrictions on the year of publication, 18 publications were obtained. Therefore, the scope was broadened as much as possible to obtain the largest number of relevant publications before applying the inclusion and exclusion criteria.

The following steps were used to run the bibliographic survey:

1. The following bibliographical databases were selected to carry out the queries:
   - ACM Digital Library, available at https://dl.acm.org.
   - Directory of Open Access Journals, available at https://doaj.org.
   - IEEE Xplore, available at https://ieeexplore.ieee.org.
   - OpenAlex, available at https://openalex.org.
   - Open Knowledge Maps, available at https://openknowledgemaps.org.
   - Science Direct, available at https://www.sciencedirect.com.
   - Scopus, available at https://www.scopus.com.
   - Web of Science, available at https://www.webofscience.com.
2. Query string: (“`physics informed`” `OR` “`physics based`”) `AND` “`solar flare`”.
3. The following inclusion and exclusion criteria were defined for the materials collected in the queries:
   - Inclusion Criteria:
     - Studies that use physics-informed machine learning models for solar flare forecasting.
     - Studies that present a clear research methodology.
     - Studies that present conclusions based on data.
   - Exclusion criteria:
     - Studies not published in English.
     - Studies that did not attend to the inclusion criteria.
     - Books or gray literature[1].

Section 5.2 presents the quantitative results of the bibliographic survey.

5.2 Quantitative results of the bibliographic survey

Figure 9 illustrates the results of the literature survey. Using the query string, the eight bibliographic databases indicated in Section 5.1 were searched. The initial search returned 39 records, of which six were duplicates.

Bibliographical databases with a more general scope and indexed publications (Scopus and Web of Science, highlighted in green) yielded more results. Other databases with a general scope but concentrating on open-access publications (highlighted in orange) yielded few articles (two per database). Finally, the bibliographic databases with a more restricted scope to the areas of Computing and Engineering (highlighted in blue) resulted in few articles (one and three, respectively).

In a quick initial analysis, the bibliographic survey conducted shows that the application of physics-informed machine learning for solar flare forecasting is in its early stages, with few publications.

The relevance of the publications was then assessed by analyzing each publication's title and summary. At this stage, of the 33 remaining records ($39 - 6$ duplicates), 17 were removed because the title,

[1] Grey literature refers to documents produced by non-scientific publishing channels or without peer review, such as reports, white papers, preprints, opinion pieces, etc.

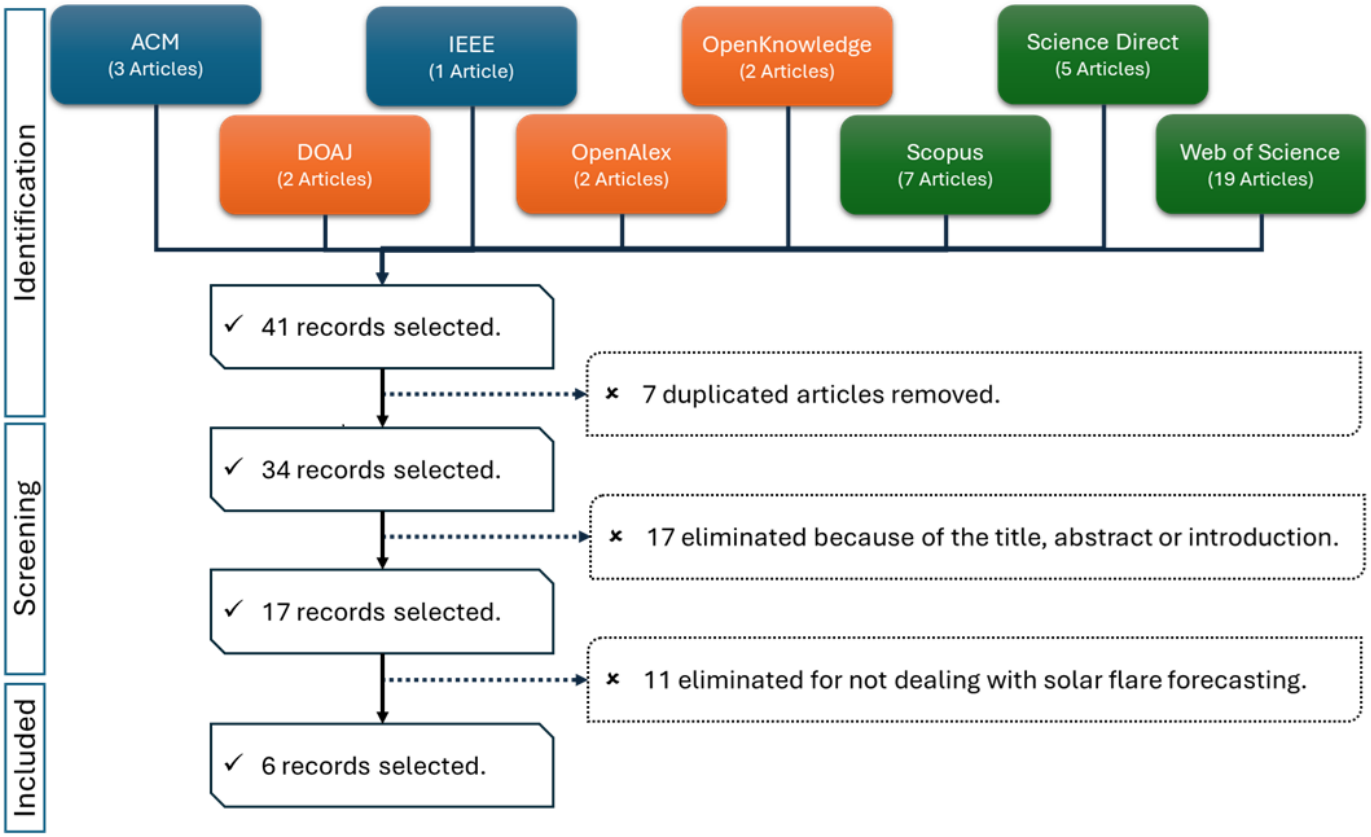

Figure 9. Results from bibliographic survey.

abstract, or introduction indicated that the works addressed topics other than the ones being sought.

Next, after analyzing the remaining 16 records, 11 publications were eliminated because they did not directly or indirectly deal with solar flare forecasting. In the end, six records [2, 20, 28, 36, 41, 56] remained.

5.3 Quantitative analysis from the bibliographic survey

Based on the bibliographic survey and a preliminary analysis of the content of the collected material, the selected articles were read more in-depth, which is discussed next.

First, Aktukmak et al. [2] explore using the Sun's north and south polar field intensities to improve solar flare forecasting using machine learning models. When combined with active regions' local data, the polar field data provides information global to the predictor. The authors propose a probabilistic mixture-of-experts (MoE) model that effectively incorporates polar field data and achieves performance comparable to leading machine learning algorithms. Their experimental results indicated that using polar field data for solar flare forecasting improved the HSS metric by 10.1%.

In turn, Li et al. [36] focused on developing deep neural network models based on knowledge for solar flare forecasting. The authors integrate prior knowledge about flare production in a convolutional neural network (CNN) structure. This knowledge guides model training in data processing, sample clustering, and implementation of additional input parameters. The results indicate that incorporating prior knowledge improves the performance of deep learning eruption prediction models (43% increase on HSS and 23% in TSS).

In this sense, magnetic field measurements are a primary source of information. Thus, photospheric vector magnetograms, obtained by instruments such as the Helioseismic and Magnetic Imager (HMI) on board the Solar Dynamics Observatory (SDO), provide important data on the intensity and direction of magnetic fields in active solar regions. These data are essential for understanding the buildup and sudden release of magnetic energy that drives solar flares.

In Li et al. [36] work, the authors integrate prior knowledge in three main aspects:

- Data preprocessing: keep the magnetic structure unchanged when resizing the active regions (ARs) magnetograms, filling them into square images.
- Sample grouping: Split the data into subsets based on different flares productivity (e.g., ARs with simple magnetic types or small areas have low productivity).
- Implementation of extra input parameter: To add the AR area as an additional input parameter to the convolutional neural network.

Solar flare productivity was also correlated with the area of the AR. Larger areas correlated with higher productivity. Statistical analyses confirmed that ARs with simple magnetic types (Alpha) or small areas ($<$ 100 millionths) rarely produce large flares, justifying the grouping of samples.

In another work, Liu et al. [41] present a hierarchy of machine learning models to forecast solar irradiances, including global horizontal irradiance, direct normal irradiance, and diffuse horizontal irradiance. The models progressively incorporate additional physics-based predictors, such as cloud fraction, cloud albedo, and clear-sky irradiances. The authors compare different machine learning models to physics-informed persistence models by evaluating their forecasting accuracy. The study notes the limited use of interpretability approaches and the underlying physical principles in many machine learning models.

In turn, Sun et al. [56] address the classification of solar flares using machine learning models. The study aims to overcome the limitations of traditionally employed predictors. The work acknowledges that the SHARP parameters [9] are widely employed. However, these parameters are scalar quantities that summarize physical information through averages or spatial integration. As a result, they lose the two-dimensional spatial distribution of the magnetic field and related quantities.

To improve accuracy, Sun et al. [56] propose creating new features that capture the local spatial distribution of the magnetic field, particularly in the Polarity Inversion Line (PIL) region. A PIL is a line on the Sun's surface that separates regions of opposite magnetic polarity. These lines are often found beneath solar filaments and are critical areas where magnetic energy builds up. They are sites of intense solar activity, like flares and coronal mass ejections. Strong PILs can act as a shear layer where the magnetic field changes direction.

The main differential in the work [56] lies in the introduction of two new sets of features that complement the SHARP parameters. The first set uses Topological Data Analysis (TDA), specifically the persistence homology, to summarize the geometric information of the distribution of various SHARP quantities. For example, it quantifies the number of loops formed by high-flux pixels. The second set employs spatial statistics tools, such as Ripley's K function and the variogram (a description of the spatial continuity of the data), to analyze the vertical magnetic field component (B_r).

Ripley's K function analyzes the clustering and dispersion pattern of high-B_r pixels at various scales. The variogram summarizes the spatial variation of the B_r flux at different distances. Sun et al.[56] restricted all feature construction to the region masked by the PIL, which is the epicenter of flare activity. This limitation helps remove noise and focuses on the most relevant regions.

The results show that using the new features, either alone or with the SHARP parameters, can significantly improve the TSS of the flare classification model [56]. An important finding is that the spatial statistics features, especially Ripley's K function, have a higher individual feature importance than the SHARP parameters. Sun et al. [56] also suggest that the Br component alone can generate powerful predictors. Specifically, M-flares tend to exhibit a B_r pattern with small PIL areas and large clusters of high-B_r pixels close to each other. There is also a large variation in B_r value between clusters. This pattern suggests that the degree of concentration of high-flux regions along the PIL is a crucial discriminator.

In turn, Guastavino et al. [20] address the prediction of travel times for Coronal Mass Ejections (CMEs) from the Sun to Earth. CMEs, massive eruptions of plasma and magnetic field, propagate through the heliosphere, the region influenced by the solar wind and the Sun's magnetic field. They are scientifically significant and, in space weather, are correlated with geomagnetic storms (disturbances in Earth's magnetic field) and solar flares.

The study introduces a physics-driven artificial intelligence (AI) approach, combining the computational efficiency of AI with the physical information in deterministic models. The authors utilize the Drag-Based Model (DBM), which estimates drag forces, to enhance training in a cascaded array of

two neural networks, where the output of one network serves as the input to the next.

The DBM is widely used due to its computational efficiency and the limited number of input parameters, describing the kinematics of the CME through its interaction with the solar wind. The first network (N1) is used to estimate the drag parameter (C), which is the only DBM parameter not provided by experimental measurements. The second network (N2) uses the results of N1, along with remote sensing and *in situ* data (such as initial CME velocity and solar wind speed), to predict travel time. Physical information is incorporated through loss functions, where N1 uses a loss function entirely inspired by the model, and N2 uses a weighted sum of data-driven and model-driven components.

The results highlight that incorporating physics-based constraints into the AI architecture substantially improves travel time predictions for CMEs compared to a purely data-driven approach. The most accurate and robust predictions are obtained when combining data-driven and physics-driven loss functions and explicitly using the drag parameter (C) as input for the second network.

Permutation Importance analysis found drag parameter, wind speed, and initial CME velocity most strongly impact predictions. Including C as a feature improved forecast robustness and reduced the number of outliers. The worst performance was with Configuration 1, which is fully data-driven.

On the other front, Jarolim et al. [28] present a new approach to extrapolating the coronal magnetic field using physics-informed neural networks. The method integrates observational data and the physical model of the Nonlinear Force-Free field (NLFF). It allows for the flexible balance between observation and the hypothesis of absence of strength, improving the understanding of the connection between observation and the underlying physics.

The method proposed by Jarolim et al. [28] uses physics-informed neural networks to extrapolate the coronal magnetic field from photospheric data. They optimized the neural network to match frontier data and satisfy physical equations (NLFF and divergence zero). The authors showed a tradeoff between the model's data agreement and the underlying Physics. Applying meta-learning (using previous results as a starting point) allows for much faster time series simulations.

It is essential to note that Jarolim et al. [28] are relevant to this section because they demonstrate how integrating observational data with a physics-based force-free magnetic-field model, facilitated by a neural network, can enhance coronal magnetic-field extrapolation, a method potentially applicable to solar flare forecasting.

Table 2 provides a comparative summary of the works discussed in this section, highlighting the relationships between different studies and solar flare forecasting, as well as other solar phenomena, irradiance prediction [41], coronal mass ejection [20] or coronal field reconstruction [28]. The primary purpose of Table 2 is to illustrate how insights and physics-informed methods from these broader solar applications can guide and enhance the development of more accurate and robust solar flare forecasting techniques.

6 Final remarks

Despite a few publications up to date, the use of physics-informed machine learning models has considerable potential to carry out solar flare forecasting with higher quality. Such superior quality should not be measured exclusively by higher precision or broader forecasting horizons, but also by other criteria. For example, the greater capacity for interpretation – that is, the degree to which a human being can understand the reasoning behind the predictions or decisions of the model –, the greater power of generalization, and the greater robustness of the model.

Table 3 compares the pure data-driven methods with physics-informed machine learning methods for solar flare forecasting.

It is important to highlight that transitioning from purely data-driven models to hybrid models (data plus scientific knowledge) can lead to discoveries in solar flare forecasting, with a deeper physical understanding of the complex processes that lead to solar flares compared to traditional prediction approaches.

However, it is important to recognize that the challenges related to the shortage of data, particularly those on high-intensity solar flares, the inherent complexity of solar phenomena, and the rigorous validation of these advanced models remain significant areas of ongoing focus and research.

Looking further, the potential for advances in space weather prediction using PIML is substantial. The continued development of more sophisticated PIML

Table 2. Comparative analysis of selected work on Physics-Informed Machine Learning in space weather.

Work	Primary Objective	Key Model	Knowledge Integration Approach	Key Findings
[2]	Improving solar flare prediction by incorporating global polar field data.	Comparison of Logistic Regression, Mixture of Experts (MoE), MLP, and Recurrent Neural Network (RNN).	Incorporates global solar cycle physics via polar field strength, which correlates with solar cycle intensity and AR activity.	Polar field data improved performance significantly in multiple configurations. RNN showed the largest mean improvement, boosting the HSS by as much as 10.1%. MOE is a simple yet effective alternative.
[36]	Forecasting ($\geq$ M-class) solar flares in the next 48 hours.	Knowledge-Informed Deep Neural Networks built upon CNN structures. The Fusion Model 2 (F2) incorporates the most prior knowledge.	Prior knowledge of flare production guides data preprocessing, sample grouping, and the addition of an AR area as an extra input parameter.	Fusion Model 2 achieved the highest TSS and F1 score as the best forecasting performance among the models tested.
[41]	Forecasting Solar irradiances, aimed at enhancing model interpretability.	Comparison of ARIMA, LSTM, and XGBoost models within a physics-based hierarchical framework.	A hierarchy of ML models is developed, where predictors are added step-by-step based on known cloud-radiation theoretical relationships. ML performance is compared against physics-informed persistence models.	LSTM and XGBoost outperform all persistence models at most lead times. Adding cloud radiation properties resulted in a significant improvement.
[56]	Improving and interpreting solar flare predictions.	XGBoost model using newly constructed Topological Features and Spatial Statistics Features (SSF).	Features capture spatial complexity and clustering patterns near the Polarity Inversion Line, relating structure to the buildup of free energy.	The constructed SSF showed the highest individual discriminating power. Combining all features improved the TSS compared to using SHARP parameters only.
[20]	Prediction of Coronal Mass Ejections' travel times.	Physics-driven AI approach using a cascade of two Neural Networks.	The deterministic Drag-Based Model (DBM) is exploited to improve the training phase. The DBM equation is encoded in the loss function.	The use of physical information improves prediction accuracy and robustness.
[28]	Probing/extrapolation of the solar coronal magnetic field.	Physics-Informed Neural Networks act as a function mapping coordinates to the magnetic field vector.	Physical Nonlinear Force-Free (NLFF) magnetic field equations are integrated as constraints into the loss function.	Enables *quasi* real-time NLFF extrapolations. Depletion of free magnetic energy aligns unambiguously with observed flare activity.

models capable of capturing the complex physics of solar flares with greater fidelity is expected.

Additionally, advances in computing power will further enable the use of complex simulations and the training of deep machine learning with bigger and more comprehensive datasets. Improvements in the availability and quality of solar observational data, including magnetic vector field measurements and high-resolution images in multiple wavelengths, will provide richer inputs to those models. The development of standardized assessment metrics and robust validation structures will be crucial to objectively assess the performance and reliability of forecasting systems based on PIML.

Integrating PIML techniques with other domains of space weather forecasting, such as coronal mass ejections and solar particle events forecasting, could lead to a more holistic and comprehensive space weather forecasting system. Furthermore, exploring explainable artificial intelligence techniques within the PIML framework will be vital to improving the interpretability of these complex models, building greater trust among astrophysicists and end users, and enabling new scientific discoveries about the Sun.

Table 3. Comparison of Pure Data-Driven and Physics-Informed Machine Learning Methods for solar flare forecasting.

Feature	Pure Data Driven Methods	Physics-Informed ML Methods
Modeling Approach	Statistical models, empirical relationships, expert systems.	Machine learning models integrated with physical constraints.
Interpretability	Generally interpretable based on known physical proxies.	It can be challenging for purely data-driven ML, but PIML aims to improve.
Ability to Capture Physics	Indirectly through proxies like sunspots.	Directly integrates physical laws and principles.
Temporal Dynamics	May overlook the evolutionary nature of solar activity.	Can model temporal sequences and dependencies.
Generalization	Performance can vary across solar cycles.	Potential for better generalization due to physical constraints.
Handling Complexity	Struggles with highly non-linear and multi-faceted relationships.	Better equipped to handle complex interactions and high-dimensional data.
Data Used	Sunspot observations, magnetic field classifications, historical flare rates.	Magnetic field measurements (photospheric, coronal), solar activity indices, EUV/X-ray data, physical laws.
Data Efficiency	Can require significant historical data.	Potential for better performance with limited or noisy data.
Feature Extraction	Often manual, based on established physical parameters.	It can be automated through deep learning, with physics guiding the process.
Accuracy	Limited, struggles with precise timing and intensity.	Potential for improved accuracy and longer lead times.
Computational Cost	Generally lower.	Can range from moderate to high, depending on model complexity.

6.1 Transitioning Space Weather Forecast Models from Research to Operations

The primary challenge in integrating machine learning models into operational space weather forecasting is ensuring that robust evaluation and practical implementation directly support actionable, reliable forecasts that inform decision-makers. Meeting this need is central to operational forecasting.

To deliver actionable operational forecasts, a robust evaluation of machine learning models is essential. Variation in forecasting performance, often caused by dataset construction, means that reliable and representative training and validation sets are crucial for operational success, specifically those accounting for solar cycle periodicity.

For operational viability, training data must closely match the forecasting context. In solar flare forecasting, this means training sets should reflect flare class rates for the relevant solar cycle phase. Precise binary definitions such as flare/no-flare using peak flux thresholds, are essential, as optimal features vary by event class and guide the transition from research to operations.

Achieving operational excellence requires a unified focus on standardizing validation and institutional support. Participation from NASA's Community Coordinated Modeling Center (CCMC), NOAA

Space Weather Prediction Center, Met Office Space Weather Operations Centre, Brazilian Studies and Monitoring of Space Weather, the community-driven COSPAR/International Space Weather Action Teams initiative, and international partners is crucial to maintaining and raising standards. Ultimately, centering efforts on delivering reliable, tailored operational forecasts for stakeholders drives advancement in space weather forecasting.

6.2 Research proposals for PIML-based solar flare forecasting

To move from data-driven methods to more robust Physics-Informed Machine Learning (PIML) models, some actions are proposed. Table 4 lists research actions for PIML-based solar flare forecasting, organized into short, mid, and long-term goals. This roadmap outlines a path from foundational data and feature enhancements to advanced hybrid models, leading to the ultimate goal of model generalizability and new physical discoveries.

Table 4. Research proposals for PIML-based solar flare forecasting in the short, mid, and long-term.

Timeframe	Research Actions & Focus
Short-Term	**Data & Feature Enhancement** • Integrating diverse data sources (e.g., local AR data with global-scale data). • Improving feature engineering to move beyond simple scalar summaries (like standard SHARP parameters). • Developing new features that capture the 2D spatial distribution of magnetic fields (especially around the PIL). • Applying advanced statistical or topological methods (TDA, Ripley's K function) to extract more physically meaningful information.
Mid-Term	**Hybrid Model & Architecture Development** • Embedding known physical laws, conservation laws, and PDEs directly into the model's loss function as a strong regularizer. • Designing specialized neural network architectures inherently structured to respect physical principles. • Quantifying the trade-off between the computational cost of PIML models and the measurable improvements in performance metrics. • Using PIML to improve model performance in scenarios with sparse, noisy, or incomplete data.
Long-Term	**Physics Discovery & Generalizability** • Enhancing interpretability and explainability to move beyond "black box" forecasting and gain new insights. • Using PIML to solve inverse problems: inferring hidden physical parameters from observational data. • Applying models to discover hidden physical relations in magnetic energy storage and release processes. • Achieving true model robustness and generalizability, ensuring physical plausibility and adaptation to new data.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The author declares no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

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

[2] Aktukmak, M., Sun, Z., Bobra, M., Gombosi, T., Manchester IV, W. B., Chen, Y., & Hero, A. (2022). Incorporating polar field data for improved solar flare prediction. *Frontiers in Astronomy and Space Sciences, 9*, 1040107. [CrossRef]

[3] Aliferis, C., & Simon, G. (2024). Overfitting, underfitting and general model overconfidence and under-performance pitfalls and best practices in machine learning and AI. *Artificial intelligence and machine learning in health care and medical sciences: Best practices and pitfalls*, 477-524. [CrossRef]

[4] Antiochos, S. K., DeVore, C. R., & Klimchuk, J. A. (1999). A model for solar coronal mass ejections. *The Astrophysical Journal, 510*(1), 485–493. [CrossRef]

[5] Asensio Ramos, A., Cheung, M. C. M., Chifu, I., & Gafeira, R. (2023). Machine learning in solar physics. *Living Reviews in Solar Physics, 20*(1), 4. [CrossRef]

[6] Azari, A. R., Lockhart, J. W., Liemohn, M. W., & Jia, X. (2020). Incorporating physical knowledge into machine learning for planetary space physics. *Frontiers in Astronomy and Space Sciences, 7*, 36. [CrossRef]

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

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

[9] Bobra, M. G., Sun, X., Hoeksema, J. T., Turmon, M., Liu, Y., Hayashi, K., Barnes, G., & Leka, K. D. (2014). The helioseismic and magnetic imager (hmi) vector magnetic field pipeline: Sharps – space-weather hmi active region patches. *Solar Physics, 289*(9), 3549–3578. [CrossRef]

[10] Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather, 17*(8), 1166–1207. [CrossRef]

[11] Carmichael, H. (1964). on the Physics of Solar Flares. In *Proc. of AAS-NASA Symp.* (Vol. 451). NASA Spec. Pub..

[12] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. [CrossRef]

[13] Doswell, C. A., Davies-Jones, R., & Keller, D. L. (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting, 5*(4), 576–585. [CrossRef]

[14] Farea, A., Yli-Harja, O., & Emmert-Streib, F. (2024). Understanding physics-informed neural networks: Techniques, applications, trends, and challenges. *AI, 5*(3), 1534–1557. [CrossRef]

[15] Florios, K., Kontogiannis, I., Park, S. H., Guerra, J. A., Benvenuto, F., Bloomfield, D. S., & Georgoulis, M. K. (2018). Forecasting solar flares using magnetogram-based predictors and machine learning. *Solar Physics, 293*(2), 28. [CrossRef]

[16] Francisco, G., Berretti, M., Chierichini, S., Mugatwala, R., Fernandes, J., Barata, T., & Del Moro, D. (2025). Limits of solar flare forecasting models and new deep learning approach. *The Astrophysical Journal, 985*(1), 108. [CrossRef]

[17] Goodwin, G. T., Sadykov, V. M., & Martens, P. C. (2024). Investigating performance trends of simulated real-time solar flare predictions: The impacts of training windows, data volumes, and the solar cycle. *The Astrophysical Journal, 964*(2), 163. [CrossRef]

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

[19] Guerra, J. A., Murray, S. A., Bloomfield, D. S., & Gallagher, P. T. (2020). Ensemble forecasting of major solar flares: methods for combining models. *Journal of Space Weather and Space Climate, 10*, 38. [CrossRef]

[20] Guastavino, S., Candiani, V., Bemporad, A., Marchetti, F., Benvenuto, F., Massone, A. M., Mancuso, S., Susino, R., Telloni, D., Fineschi, S., & Piana, M. (2023). Physics-driven machine learning for the prediction of coronal mass ejections' travel times. *The Astrophysical Journal, 954*(2), 151. [CrossRef]

[21] Guastavino, S., Marchetti, F., Benvenuto, F., Campi, C., & Piana, M. (2023). Operational solar flare forecasting via video-based deep learning. *Frontiers in Astronomy and Space Sciences, 9*, 1039805. [CrossRef]

[22] Hanslmeier, A. (2010). The sun and space weather. In *Heliophysical Processes* (pp. 233-249). Berlin, Heidelberg: Springer Berlin Heidelberg. [CrossRef]

[23] Hao, Z., Liu, S., Zhang, Y., Ying, C., Feng, Y., Su, H., & Zhu, J. (2022). Physics-informed machine learning: A survey on problems, methods and applications. *arXiv preprint arXiv:2211.08064*.

[24] Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer, 27*(2), 83-85.

[25] Heyvaerts, J., Priest, E. R., & Rust, D. M. (1977). An emerging flux model for the solar flare phenomenon. *Astrophysical Journal, Part 1, vol. 216, Aug. 15, 1977, p. 123-137., 216*, 123-137. [CrossRef]

[26] Hirayama, T. (1974). Theoretical model of flares and prominences: I: Evaporating flare model. *Solar Physics, 34*(2), 323–338. [CrossRef]

[27] Hirose, S., Uchida, Y., Uemura, S., Yamaguchi, T., & Cable, S. B. (2001). A Quadruple Magnetic Source Model for Arcade Flares and X-RayArcade Formations outside Active Regions. II. Dark FilamentEruption and the AssociatedArcade Flare. *The Astrophysical Journal, 551*(1), 586. [CrossRef]

[28] Jarolim, R., Thalmann, J. K., Veronig, A. M., & Podladchikova, T. (2023). Probing the solar coronal magnetic field with physics-informed neural networks. *Nature Astronomy, 7*(10), 1171–1179. [CrossRef]

[29] Ji, A., Aydin, B., Georgoulis, M. K., & Angryk, R. (2020). All-clear flare prediction using interval-based time series classifiers. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 4218–4225). IEEE. [CrossRef]

[30] Jiao, Z., Sun, H., Wang, X., Manchester, W., Gombosi, T., Hero, A., & Chen, Y. (2020). Solar flare intensity prediction with machine learning models. *Space weather, 18*(7), e2020SW002440. [CrossRef]

[31] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics, 3*(6), 422–440. [CrossRef]

[32] Kopp, R. A., & Pneuman, G. W. (1976). Magnetic reconnection in the corona and the loop prominence phenomenon. *Solar Physics, 50*(1), 85-98. [CrossRef]

[33] Kusano, K., Iju, T., Bamba, Y., & Inoue, S. (2020). A physics-based method that can predict imminent large solar flares. *Science, 369*(6503), 587-591. [CrossRef]

[34] Lam, H. L., & Samson, J. C. (1994). An investigation of the time-delay between solar events and geomagnetic disturbances using a new method of superposed epoch analysis. *Journal of geomagnetism and geoelectricity, 46*(2), 107-113. [CrossRef]

[35] Leka, K. D., Park, S. H., Kusano, K., Andries, J., Barnes, G., Bingham, S., ... & Terkildsen, M. (2019). A comparison of flare forecasting methods. II. Benchmarks, metrics, and performance results for operational solar flare forecasting systems. *The Astrophysical Journal Supplement Series, 243*(2), 36. [CrossRef]

[36] Li, M., Cui, Y., Luo, B., Ao, X., Liu, S., Wang, J., ... & Wang, X. (2022). Knowledge-informed deep neural networks for solar flare forecasting. *Space weather, 20*(8), e2021SW002985. [CrossRef]

[37] Li, R., & Zhu, J. (2013). Solar flare forecasting based on sequential sunspot data. *Research in Astronomy and Astrophysics, 13*(9), 1118–1126. [CrossRef]

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

[39] Liu, C., Deng, N., Wang, J. T. L., & Wang, H. (2017). Predicting solar flares using SDO/HMI vector magnetic data products and the random forest algorithm. *The Astrophysical Journal, 843*(2), 104. [CrossRef]

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

[41] Liu, W., Liu, Y., Zhang, T., Han, Y., Zhou, X., Xie, Y., & Yoo, S. (2022). Use of physics to improve solar forecast: Part ii, machine learning and model interpretability. *Solar Energy, 244*, 362–378. [CrossRef]

[42] Liu, W., Liu, Y., Zhou, X., Xie, Y., Han, Y., Yoo, S., & Sengupta, M. (2021). Use of physics to improve solar forecast: Physics-informed persistence models for simultaneously forecasting GHI, DNI, and DHI. *Solar Energy, 215*, 252-265. [CrossRef]

[43] Lysenko, A. L., Frederiks, D. D., Fleishman, G. D., Aptekar, R. L., Altyntsev, A. T., Golenetskii, S. V., Svinkin, D. S., Ulanov, M., Tsvetkova, A. E., & Ridnaia, A. V. (2020). X-ray and gamma-ray emission from solar flares. *Physics-Uspekhi, 63*(8), 818–832. [CrossRef]

[44] Messerotti, M., Zuccarello, F., Guglielmino, S. L., Bothmer, V., Lilensten, J., Noci, G., ... & Lundstedt, H. (2009). Solar weather event modelling and prediction. *Space science reviews, 147*(3), 121-185. [CrossRef]

[45] Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning* (p. 78). [CrossRef]

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

[47] Otto, P., Fassò, A., & Maranzano, P. (2024). A review of regularised estimation methods and cross-validation in spatiotemporal statistics. *Statistic Surveys, 18*, 299-340. [CrossRef]

[48] Pateras, J., Rana, P., & Ghosh, P. (2023). A taxonomic survey of physics-informed machine learning. *Applied Sciences, 13*(12), 6892. [CrossRef]

[49] Popper, K. R. (2010). *The logic of scientific discovery* (Special Indian edition). Routledge.

[50] Priest, E. R., Parnell, C. E., & Martin, S. F. (1994). A converging flux model of an x-ray bright point and an associated canceling magnetic feature. *The Astrophysical Journal, 427*, 459. [CrossRef]

[51] Quarteroni, A., Gervasio, P., & Regazzoni, F. (2025). Combining physics-based and data-driven models: advancing the frontiers of research with scientific machine learning. *arXiv preprint arXiv:2501.18708*.

[52] Ribeiro, F., & Gradvohl, A. L. S. (2021). Machine learning techniques applied to solar flares forecasting. *Astronomy and Computing, 35*, 100468. [CrossRef]

[53] Seyyedi, A., Bohlouli, M., & Oskoee, S. N. (2024). Machine learning and physics: A survey of integrated models. *ACM Computing Surveys, 56*(5), 1–33. [CrossRef]

[54] Stankewitz, B. (2024). Early stopping for L 2-boosting in high-dimensional linear models. *The Annals of Statistics, 52*(2), 491-518. [CrossRef]

[55] Sturrock, P. A. (1966). Model of the high-energy phase of solar flares. *Nature, 211*(5050), 695–697. [CrossRef]

[56] Sun, H., Manchester IV, W., & Chen, Y. (2021). Improved and interpretable solar flare predictions with spatial and topological features of the polarity inversion line masked magnetograms. *Space weather, 19*(12), e2021SW002837. [CrossRef]

[57] Tang, R., Liao, W., Chen, Z., Zeng, X., Wang, J. S., Luo, B., ... & Wu, Z. (2021). Solar flare prediction based on the fusion of multiple deep-learning models. *The Astrophysical Journal Supplement Series, 257*(2), 50. [CrossRef]

[58] Vural, O., Hamdi, S. M., & Boubrahimi, S. F. (2025). Solar Flare Prediction Using Multivariate Time Series of Photospheric Magnetic Field Parameters: A

Comparative Analysis of Vector, Time Series, and Graph Data Representations. *Remote Sensing*, *17*(6), 1075. [CrossRef]

[59] Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. (2013, May). Regularization of neural networks using dropconnect. In *International conference on machine learning* (pp. 1058-1066). PMLR.

[60] Watson, J., Song, C., Weeger, O., Gruner, T., Le, A. T., Pompetzki, K., ... & Hoffman, M. W. (2024). Machine learning with physics knowledge for prediction: A survey. *arXiv preprint arXiv:2408.09840*.

[61] Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2022). Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, *55*(4), 1-37. [CrossRef]

André Leon S. Gradvohl received a B.S. in Computer Science from the Universidade Federal do Ceará, Brazil (1997), and an M.Sc. in Electrical Engineering and Computer Science from the Technological Institute of Aeronautics, Brazil (2000). He completed a Ph.D. in Electrical and Computer Engineering at Universidade Estadual de Campinas (UNICAMP) in 2005. In 2010, he earned a Specialist degree in Science Journalism from UNICAMP's Advanced Studies in Journalism Laboratory. He is currently an associate professor at the Faculdade de Tecnologia, UNICAMP. His research interests include High-Performance Computing, Distributed Computing, Information Security, and Solar Flare Forecasting. He is also a member of the High-Performance Intelligent Decision Systems research group. (Email: gradvohl@unicamp.br)