Flare Transformer: Solar Flare Prediction using Magnetograms and Sunspot Physical Features

Kanta Kaneda[1], Yuiga Wada[1], Tsumugi Iida[1],
Naoto Nishizuka[2], Yûki Kubo[2], and Komei Sugiura[1]

[1] Keio University, Japan
{k.kaneda, yuiga, tiida, komei.sugiura}@keio.jp
[2] National Institute of Information and Communications Technology, Japan
{nishizuka.naoto, kubo}@nict.go.jp

Abstract. The prediction of solar flares is essential for reducing the potential damage to social infrastructures that are vital to society. However, predicting solar flares accurately is a very challenging task. Existing methods predict flares using either physical features or images, but the main bottleneck is that they sometimes incorrectly predict a class that is smaller than the actual solar flare. In this paper, we propose the Flare Transformer, a solar flare prediction model that handles both images and physical features through the Magnetogram Module and the Sunspot Feature Module. The transformer attention mechanism is introduced to model the temporal relationships between input features. We also introduce a new differentiable loss function to balance the two major metrics of the Gandin–Murphy–Gerrity score and Brier skill score. We validate our model on a publicly available dataset. The results show that the Flare Transformer outperformed the baseline methods in terms of the Gandin–Murphy–Gerrity score and true skill statistic, and achieved better performance than those given by human experts.

Keywords: Solar flare prediction · Time-series forecasting · Transformer

1 Introduction

X-ray emissions, high energy particles, and coronal mass ejections released by solar flares disrupts GPS communication, causes radio blackouts, and poses health hazards to astronauts and flight crews [2]. It is estimated that the economic loss from a Carrington-class flare will be approximately US $163 billion in North America [22]. Therefore, the prediction of solar flares is essential for reducing the potential damage to our society. However, accurate predictions of solar flares remain a challenging task.

Given this background, this paper focuses on predicting the class of the largest solar flare that will occur within 24 h. Fig. 1 shows an overview of our method. The inputs are a time-series of full-disk line-of-sight magnetograms taken by Helioseismic and Magnetic Imager (HMI) [24] onboard Solar Dynamic Observatory (SDO) [20] and region-level physical features extracted from active

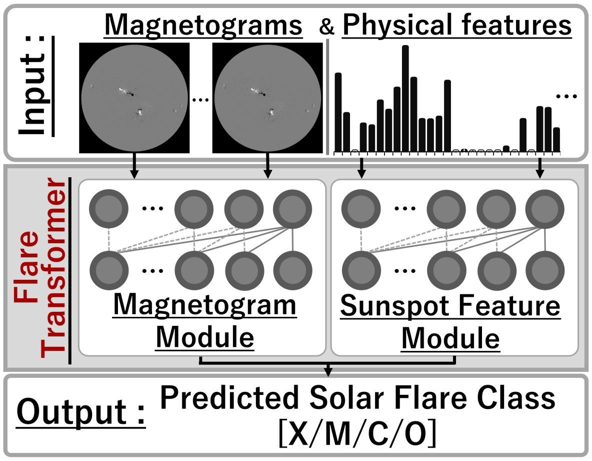

Fig. 1. Overview of the Flare Transformer.

regions detected in the sunspot images. Note that a magnetogram is a grayscale solar image shown in Fig. 1, and details on physical features can be found in [6]. The output is the predicted solar flare class (see Table 1 for details).

Even for human experts, it is very challenging to predict solar flares. For example, their performance during the period 2000–2015 resulted in GMGS = 0.48 (Gandin–Murphy–Gerrity score) and $\mathrm{TSS}_{\geq \mathrm{M}} = 0.50$ (true skill statistics) [10] (these metrics would return a score of 1 for a perfect forecast). The main bottleneck of existing methods is that they sometimes incorrectly predict a class of flare that is smaller than that which actually occurs. For example, the DeFN [17] model incorrectly predicted 89% of X-class flares as M-class.

In this paper, we propose the Flare Transformer (FT), which handles time-series of images and physical features to produce accurate and reliable solar flare predictions. Our code is available at this URL[3]. Our model differs from existing methods in that it handles both line-of-sight magnetograms and physical features with the Magnetogram Module (MM) and the Sunspot Feature Module (SFM) (see Section 4.3). These modules allow the proposed FT model to capture useful features for solar flare prediction. We also introduce a transformer attention mechanism [27] to model the temporal relationships between input features. The main contributions of this paper are summarized as follows:

- We propose the FT, a solar flare prediction model that handles both line-of-sight magnetograms and physical features through the MM and SFM (see Section 4.3).
- We introduce the transformer attention mechanism [27] to model temporal relationships between input features.

[3] https://github.com/keio-smilab21/flare_transformer

- We introduce the GMGS and Brier skill score (BSS) losses to balance the two major metrics of solar flare prediction (see Section 4.4).

2 Related Work

2.1 Time-Series Forecasting

There have been many studies in the field of time-series analysis (e.g.,[29,30]). For instance, [5] is a survey paper in the field that discusses the methods, datasets, and subtasks associated with time-series analysis.

Time-series forecasting is one of the subtasks of time-series analysis. Early studies on time-series forecasting were based on statistical models (e.g.,[3,9]). More recent methods are based on deep neural networks (DNNs), which can model complicated sequential data. The existing methods using DNNs typically use recurrent neural networks (RNNs) (e.g., [18,21,23,28]) or attention mechanisms (e.g., [13,29,30]). In terms of RNN based approaches, DeepAR [23] models future probabilistic distributions using an autoregressive RNN model, whereas DPT-DRNN [18] is an enhanced RNN model with a pre-training method that uses an autoencoder, allowing the PM2.5 concentration to be predicted from environmental monitoring data.

Inspired by the success of attention mechanism in other fields (e.g., natural language processing, computer vision), many models using a transformer [27] has been proposed for time-series forecasting. However, time-series data tends to have larger sequence length L than inputs in other fields, resulting in high computational complexity. Therefore, many existing studies have proposed methods to reduce the computational complexity. For example, LogSparse Transformer [13] propose the LogSparse attention, which reduce the computational complexity to $\mathcal{O}(L \log L)$. Adversarial Sparse Transformer [29] implements a sparse attention layer by using a sparse normalization transformation, α-entmax. Informer [30] is a transformer-based model for long-sequence time-series forecasting, which introduce a ProbSparse self-attention mechanism and generative style decoder.

2.2 Solar Flare Prediction

There have also been many studies on solar flare prediction (e.g., [11,17,19]). For instance, [7] comprehensively summarizes the methods and evaluation metrics for solar flare prediction tasks.

There are several standard datasets for solar flare prediction tasks. Nishizuka *et al.* published a dataset[1] consisting of physical features for sunspots extracted from images taken by the SDO [20] and Geostationary Operational Environmental Satellite (GOES). The dataset covers the period from June 2010 to December 2015 [16]. Angryk *et al.* published a dataset consisting of physical features of active regions extracted from Spaceweather HMI Active Region Patch series, which covers the period from May 2010 to December 2018 [1].

[1] Available at https://wdc.nict.go.jp/IONO/wdc/solarflare/index.html

Table 1. The correspondence of flare class and X-ray intensity.

Flare Class	Range of X-ray intensity [W/m^2]
X	$p_t > 10^{-4}$
M	$10^{-5} < p_t \leq 10^{-4}$
C	$10^{-6} < p_t \leq 10^{-5}$
O	$p_t \leq 10^{-6}$

Many methods have been proposed for solar flare prediction tasks. For example, Park *et al.* proposed a convolutional neural network based model to forecast solar flare occurrence using solar full-disk magnetograms [19]. DeFN [17] is a residual feed-forward network model which calculates the probability of solar flares occurring in the subsequent 24 h period using sunspot physical features. Tang *et al.* compared the performance of existing solar flare prediction models, and reported that the DeFN outperformed all other methods [4,8,25] for the 2010-2015 dataset [25]. The advantage of the DeFN model is that the physical features can be analyzed to search for those that are most effective for solar flare prediction. DeFN-R [15] is an extension of DeFN. While the DeFN is optimized for deterministic prediction, DeFN-R is optimized for a probability forecast based on the observation event rate.

3 Problem Statement

In this paper, we focus on the task of predicting the class of the largest solar flare that will occur within 24 h from time t, $\boldsymbol{y}_t$, which is defined as follows:

$$\boldsymbol{y}_t = \mathrm{flareclass}(\max\{p_{t+1}, p_{t+2}, \ldots, p_{t+24}\})\,, \tag{1}$$

where p_t and flareclass($\cdot$) denote the maximum X-ray intensity within an hour of time t and the 1-of-K representation for classes X, M, C, and O, respectively. Table 1 defines each flare class in terms of p_t.

The input and output of this task are defined as follows:

– **Input:** Line-of-sight magnetograms and physical features. The physical features are extracted from solar images taken by the Atmospheric Imaging Assembly (AIA) [12] aboard the SDO [20].

– **Output:** A four-dimensional vector that denotes the predicted probabilities for each solar flare class.

In this task, it is desirable that $p(\hat{\boldsymbol{y}}_t)$ be as close as possible to $\boldsymbol{y}_t$, where $p(\hat{\boldsymbol{y}}_t)$ denotes the output of the model. However, because solar flare prediction is a class-imbalanced problem, it is unfavorable to output trivial predictions, such as predicting all flares as O-class. Thus, to avoid such predictions, it is desirable to output $p(\hat{\boldsymbol{y}}_t)$ that maximizes metrics such as GMGS [6] and BSS [15] , which are standard metrics in the field of solar flare prediction.

In this paper, we do not model the solar flare prediction task as a regression problem for the following two reasons:

- The above classification task is the standard setting in the field of solar flare prediction (e.g., [17,19]).
- Predictions by human experts are given in the above classification setting, not as a regression task.

Therefore, it is reasonable to handle this task as a classification problem. Even though we do not explicitly consider the task as a regression problem, our model can be applied to regression tasks. We also assume that solar flare prediction is based on a full-disk image and not on a region-level image. We define line-of-sight magnetogram as a solar image taken by SDO/HMI [24]. We evaluate the model by GMGS, TSS [10] and BSS (see Section 5.2). The GMGS is a multi-class evaluation metric that considers class imbalance using the GMGS score matrix [6]. The BSS is a metric to evaluate the reliability of the forecast and is used not only in the field of solar flare prediction, but also in other fields such as weather forecasting [15].

4 Proposed Method

4.1 Novelty

The proposed method is unique in that the model handles both line-of-sight magnetograms and physical features through the Magnetogram Module (MM) and the Sunspot Feature Module (SFM) for solar flare prediction. We use the transformer attention mechanism [27] in the modules to model the temporal relationships between input features. Solar flare prediction is a time series forecasting task, and a certain degree of correlation is expected between time series features. Therefore, it is considered that the transformer, which performs correlation calculations in the self-attention and cross-attention mechanisms, is suitable for this task. The main differences between our method and existing methods (e.g., [17,19]) are as follows:

- While existing methods take only physical features [17] or images [19] as input, our model handles both features through the MM and the SFM. These modules are explained in Section 4.3.
- We introduce the GMGS and BSS losses to balance the two major metrics in solar flare prediction. These losses are explained in Section 4.4.

4.2 Input

The input $\boldsymbol{x}$ is defined as follows:

$$\boldsymbol{x} = (V_{t-k+1:t}, F_{t-k+1:t}) \ , \tag{2}$$

$$V_{t-k+1:t} = (\boldsymbol{v}_{t-k+1}, \boldsymbol{v}_{t-k+2}, ..., \boldsymbol{v}_t) \ , \tag{3}$$

$$F_{t-k+1:t} = (\boldsymbol{f}_{t-k+1}, \boldsymbol{f}_{t-k+2}, ..., \boldsymbol{f}_t) \ , \tag{4}$$

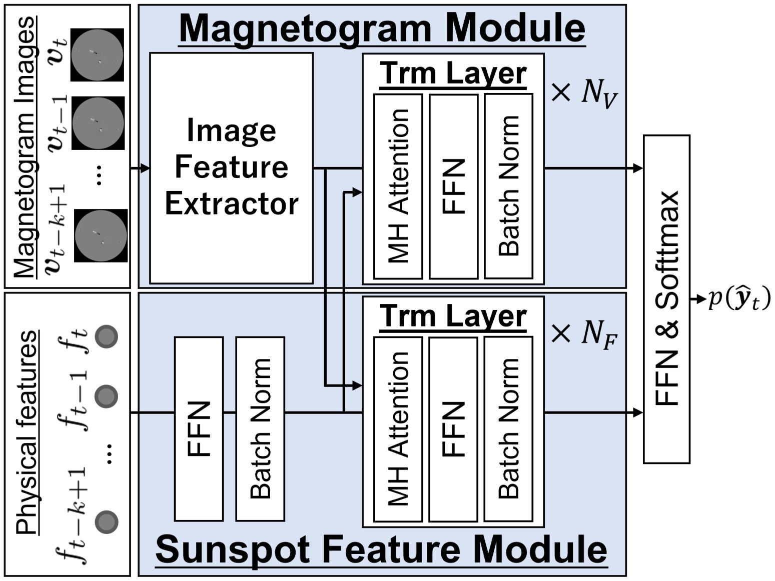

Fig. 2. Proposed method framework. Flare Transformer consists of Magnetogram Module and Sunspot Feature Module.

where $\boldsymbol{v}_t \in \mathbb{R}^{512\times512}$ and $\boldsymbol{f}_t \in \mathbb{R}^{90}$ denote the line-of-sight magnetogram and physical features at time t, respectively. We extract $\boldsymbol{f}_t$ by the method described in [17], and obtain $\boldsymbol{v}_t$ by resizing the line-of-sight magnetograms to 512×512 pixels.

4.3 Flare Transformer

Fig. 2 shows the structure of our method. In the figure, MH Attention, Trm Layer, and FFN denote the multi-head attention, transformer layer, and feed-forward network, respectively. The proposed method consists of two main modules: MM and SFM. The difference between MM and SFM is the query taken in the source-target attention. In the MM, magnetogram features are taken as the query, while in the SFM, physical features are taken as the query. The MM first encodes the images $V_{t-k+1:t}$ as follows:

$$\boldsymbol{h}_V = f_{\mathrm{FE}}\left(V_{t-k+1:t}\right) , \tag{5}$$

where f_{FE} denotes the Image Feature Extractor, which consists of multiple convolutional layers, max pooling layers, average pooling layers, and batch normalization layers, as shown in Fig. 3. The SFM also encodes the physical features $F_{t-k+1:t}$ as follows:

$$\boldsymbol{h}_F = f_{\mathrm{BN}}(f_{\mathrm{FFN}}(F_{t-k+1:t})) , \tag{6}$$

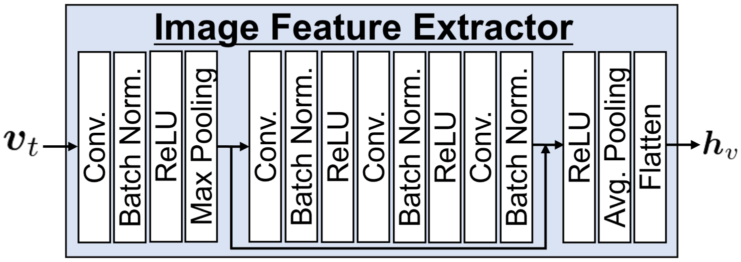

Fig. 3. The structure of Image Feature Extractor. "Conv," "Batch Norm," and "Avg pooling" denote the convolutional layer, batch normalization layer, and average pooling layer, respectively.

where f_{BN} and f_{FFN} denote the batch normalization layer and FFN, respectively. Then, we obtain $\boldsymbol{h}_{VF}$ by concatenating $\boldsymbol{h}_V$ and $\boldsymbol{h}_F$.

Next, the N_V transformer layers compute the temporal relationships between time-series images and physical features. In the multi-head attention block, $\boldsymbol{h}_V$ and $\boldsymbol{h}_{VF}$ are divided into $\boldsymbol{h}_V^{(i)} \in \mathbb{R}^{k\times d}$ and ${\boldsymbol{h}_{VF}}^{(i)} \in \mathbb{R}^{k\times 2d}\,(i = 1, \ldots, N_{\text{head}})$, where $d = H/N_{\text{head}}$. Here, H and N_{head} denote the hidden layer size and number of heads, respectively.

The query $Q^{(i)} \in \mathbb{R}^{k\times d}$, key $K^{(i)} \in \mathbb{R}^{k\times 2d}$, and value $V^{(i)} \in \mathbb{R}^{k\times 2d}$ are computed for the i-th head as follows:

$$Q^{(i)} = W_q^{(i)}\boldsymbol{h}_V^{(i)}\,, \tag{7}$$

$$K^{(i)} = W_k^{(i)}\boldsymbol{h}_{VF}^{(i)}\,, \tag{8}$$

$$V^{(i)} = W_v^{(i)}\boldsymbol{h}_{VF}^{(i)}\,, \tag{9}$$

where $W_q^{(i)}$, $W_k^{(i)}$, and $W_v^{(i)}$ denote the weight matrices for $Q^{(i)}$, $K^{(i)}$, and $V^{(i)}$, respectively. The output of the transformer layer $\boldsymbol{h}_{\text{trm}}$ is computed as follows:

$$\boldsymbol{h}_{\text{trm}} = f_{\text{BN}}(f_{\text{FFN}}(\boldsymbol{h}_{\text{mha}}))\,, \tag{10}$$

$$\boldsymbol{h}_{\text{mha}} = \left[\boldsymbol{f}_{\text{attn}}^{(1)}; \boldsymbol{f}_{\text{attn}}^{(2)}; \ldots; \boldsymbol{f}_{\text{attn}}^{(N_{\text{head}})}\right]\,, \tag{11}$$

$$\boldsymbol{f}_{\text{attn}}^{(i)} = \text{softmax}\left(\frac{Q^{(i)}K^{(i)\top}}{\sqrt{d}}\right)V^{(i)}\,. \tag{12}$$

The output of the MM $\boldsymbol{h}_{\text{MM}}$ is obtained by N_V transformer layers. Similarly, the output of the SFM $\boldsymbol{h}_{\text{SFM}}$ is obtained by N_F transformer layers.

Finally, the predicted flare class y_t^* is obtained as follows:

$$y_t^* = \text{argmax}_i(p(\hat{y}_{ti}))\,, \tag{13}$$

$$p(\hat{\boldsymbol{y}}_t) = \text{softmax}(f_{\text{FFN}}(\boldsymbol{h}_{\text{MM}}; \boldsymbol{h}_{\text{SFM}}))\,, \tag{14}$$

where $p(\hat{y}_{ti})$ denotes the predicted probability of i-th class.

4.4 Loss Function

In our model, we introduce GMGS and BSS losses to balance the GMGS [6] and $\mathrm{BSS}_{\geq \mathrm{M}}$ [15].

First, we define the GMGS loss for the following reason. The loss functions used in existing methods are not effective in improving the GMGS because they adjust the balance between classes using weights that are irrelevant to GMGS. Unlike existing methods, we can effectively improve the GMGS by using the score matrix of GMGS as weights. The GMGS loss $\mathcal{L}_{\mathrm{GMGS}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{GMGS}} = -\frac{1}{NI}\sum_{n=1}^{N} s_{i^*j^*}\sum_{i=1}^{I} y'_{ni}\log(p(\hat{y}_{ni}))\,, \tag{15}$$

$$i^* = \mathrm{argmax}_i(y_{ni})\,, \tag{16}$$

$$j^* = \mathrm{argmax}_j(p(\hat{y}_{nj}))\,, \tag{17}$$

where N, I, $p(\hat{y}_{ni})$, y_{ni}, y'_{ni} and $s_{i^*j^*}$ denote the number of samples, number of classes, predicted probability of the i-th class for the n-th sample, label of the i-th class for the n-th sample, the label of the i-th class for label smoothed $\boldsymbol{y}_n$, and element (i^*, j^*) from the score matrix for GMGS [6], respectively.

Second, we define the BSS loss to improve the reliability of the forecast. We propose to use the BSS directly for the BSS loss because it is differentiable. The BSS loss $\mathcal{L}_{\mathrm{BSS}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{BSS}} = -\frac{1}{NI}\sum_{n=1}^{N}\sum_{i=1}^{I}(p(\hat{y}_{ni}) - y_{ni})^2\,. \tag{18}$$

Overall, we use the following loss function:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{GMGS}}\mathcal{L}_{\mathrm{GMGS}} + \lambda_{\mathrm{BSS}}\mathcal{L}_{\mathrm{BSS}}\,, \tag{19}$$

where $\mathcal{L}_{\mathrm{CE}}$ denotes the cross entropy loss between $\boldsymbol{y}_n$ and $\hat{\boldsymbol{y}}_n$, and λ_{GMGS} and λ_{BSS} denote the loss weights.

5 Experiments

5.1 Experimental Setup

In the experiments, we evaluated our method on a dataset that was collected by the following procedure. First, we downloaded hourly line-of-sight magnetograms from the SDO [20] web archives[1]. Next, we used an online physical feature database that is available at this URL [2].

Because the physical features in the dataset are region-level features, the following processes were performed to make the data suitable for input to our model.

[1] https://sdo.gsfc.nasa.gov/data/

[2] https://wdc.nict.go.jp/IONO/wdc/solarflare/index.html

Table 2. The number of samples included in each divided set.

Training Set		Test Set	
Year	Samples	Year	Samples
2010-2013	29247	2014	8127
2010-2014	37374	2015	8155
2010-2015	45529	2016	7795
2010-2016	53324	2017	7991

Table 3. Parameter settings and structures of FT.

Optimizer	Adam ($\beta_1 = 0.9,\ \beta_2 = 0.999$)
Learning Rate	7.0×10^{-7}
Batch Size	32
MM	$(H_{\mathrm{MM}},\ A_{\mathrm{MM}}) = (128,\ 4)$
SFM	$(H_{\mathrm{SFM}},\ A_{\mathrm{SFM}}) = (128,\ 4)$
Trm layer	$(N_V,\ N_F) = (1,\ 2)$
Loss weights	$(\lambda_{\mathrm{GMGS}},\ \lambda_{\mathrm{BSS}}) = (0.01,\ 10)$

- If multiple sunspots were observed at a given time, we selected physical features from a randomly selected sunspot.
- If no sunspots were observed at a given time, we set the values of all physical features to zero.

The dataset contains 61315 samples, covering the period from June 2010 to December 2017. A sample consists of a line-of-sight magnetogram and 90 types of physical features [17]. The numbers of samples with ground truth labels of X, M, C, and O are 492, 4745, 19736, and 36342, respectively. The numbers of samples are imbalanced between classes because X-class and M-class solar flares are extremely unlikely events compared with other classes of flares. For example, only 2.9% of solar flares in 2017 were of X class.

In this study, we divided the training and test sets based on time-series cross-validation, which is a standard method used in time-series forecasting tasks [26]. With time-series cross-validation, the corresponding training set consists only of observations that occurred prior to the observations that form the test set. Table 2 presents the numbers of samples included in each divided set. The training and test sets were used for parameter training and evaluation, respectively.

The experimental setup is summarized in Table 3, where H_{MM} and A_{MM} denote the hidden size and number of attention heads in the MM, respectively, and H_{SFM} and A_{SFM} denote those in the SFM. Our model has 3.65 million parameters. The proposed model was trained on an RTX 2080 Ti with 11GB of GPU memory and an Intel Core i9 processor. It took approximately 90 min to train our model. The inference time was approximately 65 ms.

5.2 Evaluation Metric

We evaluate the model by GMGS [6], TSS [10] and BSS [15]. The GMGS is defined as follows:

$$\mathrm{GMGS} = \mathrm{tr}(S^\top \cdot P)\,, \tag{20}$$

where S and P denote the I-rank scoring matrix with an element s_{ij} and I-categorical contingency table with an element p_{ij}, respectively. The GMGS is used as an important metric in recent studies for solar flare prediction [10]. The elements s_{ij} for the symmetric score matrix S is defined as follows:

$$s_{ii} = \frac{1}{I-1}\left[\sum_{k=1}^{i-1} a_k^{-1} + \sum_{k=i}^{I-1} a_k\right] \quad (1 \leq i \leq I)\,, \tag{21}$$

$$s_{ij} = \frac{1}{I-1}\left[\sum_{k=1}^{i-1} a_k^{-1} + \sum_{k=i}^{j-1}(-1) + \sum_{k=j}^{I-1} a_k\right] \quad (1 \leq i \leq j \leq I)\,, \tag{22}$$

$$a_i = \frac{1-\sum_{k=1}^{i} p_k}{\sum_{k=1}^{i} p_k} \quad (1 \leq i \leq I)\,, \tag{23}$$

$$p_i = \sum_{j=1}^{I} p_{ij} \quad (1 \leq j \leq I)\,. \tag{24}$$

The TSS is defined as follows:

$$\mathrm{TSS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} - \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}, \tag{25}$$

where TP, FP, FN, TN denote the number of true positive, false positive, false negative, and true negative samples for a contingency table, respectively.

The BSS is a standard metric for solar flare prediction that evaluates the reliability of the forecast [15]. The BSS is defined as follows:

$$\mathrm{BSS} = \frac{\mathrm{BS}-\mathrm{BS}_c}{0-\mathrm{BS}_c}\,, \tag{26}$$

$$\mathrm{BS} = \sum_{n=1}^{N}\sum_{i=1}^{I}(p(\hat{y}_{ni}) - y_{ni})^2\,, \tag{27}$$

$$\mathrm{BS}_c = \sum_{n=1}^{N}\sum_{i=1}^{I}(f - y_{ni})^2\,, \tag{28}$$

where N, I, y_{ni}, $p(\hat{y}_{ni})$, and f denote the number of samples, the number of classes, the label of the i-th class for the n-th sample, the predicted probability of the i-th class for the n-th sample, and climatological event rate, respectively.

Table 4. Quantitative comparison. The best scores are in bold.

Method	GMGS↑	$\mathrm{TSS}_{\geq \mathrm{M}}$↑	$\mathrm{BSS}_{\geq \mathrm{M}}$↑
DeFN [17]	0.375±0.141	0.413±0.150	-0.022±0.782
DeFN-R [15]	0.302±0.055	0.279±0.162	0.036±0.982
Ours (FT)	**0.503±0.059**	**0.530±0.112**	**0.082±0.974**
Human [10,14]	0.48	0.50	0.16

5.3 Experimental Results

We conducted experiments based on time-series cross validation. Table 4 shows the quantitative results of the baselines and proposed method. The average and standard deviations of the scores are reported. The DeFN [17] and DeFN-R [15] models, which only take physical features as input, were used as the baseline methods. The results given by DeFN and DeFN-R were reproduced by ourselves. We evaluated the models by GMGS [6], $\mathrm{TSS}_{\geq \mathrm{M}}$ [10] and $\mathrm{BSS}_{\geq \mathrm{M}}$ [15]. "≥M" indicates that the model was evaluated after the output had been categorized as either "≥M" or "<M". We used GMGS and $\mathrm{BSS}_{\geq \mathrm{M}}$ as the primary metrics. GMGS is a metric for multi-categorical forecasts and satisfies equitability [10], whereas BSS is a standard metric for solar flare prediction that evaluates the reliability of the forecast [15]. The approximated version of GMGS is shown for the 2016 test set because no X-class flares occurred in 2016.

Table 4 indicates that the GMGS of the DeFN, DeFN-R, and FT methods are 0.375, 0.302 and 0.503 points, respectively. Therefore, the FT outperformed DeFN by 0.128 points in terms of GMGS. Table 4 also presents the performance of human experts. Kubo *et al.* reported that GMGS and $\mathrm{TSS}_{\geq \mathrm{M}}$ for daily forecasting operations by human experts were 0.48 and 0.50, respectively, for the period 2000–2015 [10]. Our method outperform human experts in terms of GMGS and $\mathrm{TSS}_{\geq \mathrm{M}}$, which indicates that FT is very promising.

Fig. 4 shows the qualitative results. Fig. 4(a) shows line-of-sight magnetograms from 21:00–23:00 on September 5, 2017 and standardized physical features. Note that, of the 90 features, only the first 20 positive values of $\boldsymbol{f}_t$ are displayed because of space limitation. The prediction is $y_t^* =$"X". An X-class solar flare occurred at 12:02 on September 6, 2017, which is within 24 h of 23:00 on September 5, 2017. Therefore, the model was able to predict the correct maximum solar flare class. Similarly, a sample that was correctly predicted as an M–class solar flare is shown in Fig. 4(b). Fig. 4(c) shows a failure case. The prediction is $y_t^* =$"M" with $p(y_t^* =$"M"$) = 0.47$ and $p(y_t^* =$ "X"$) = 0.40$. Because an X-class solar flare occurred within 24 h of time t, the prediction was incorrect. However, the predicted probability indicates that this was a marginal prediction. The above results indicate that our model gives better predictions than can be achieved by human experts in terms of GMGS and TSS.

5.4 Ablation Studies

We set the following ablation conditions :

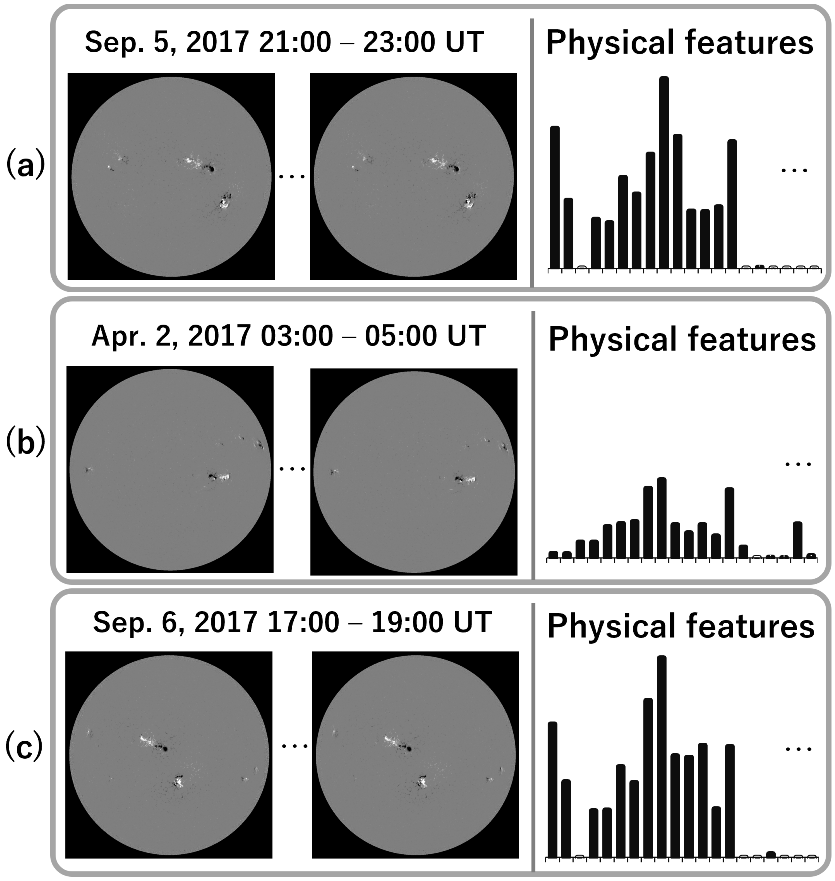

Fig. 4. Three samples of qualitative results. The model predicted the correct flare class for samples (a) and (b), while the model predicted the incorrect flare class for sample (c). (a) X-class solar flare. (b) M-class solar flare. (c) X-class solar flare but was predicted as M-class solar flare.

(a) We removed $F_{t-k+1:t}$ to investigate the performance when $V_{t-k+1:t}$ is used as the input.

(b) We changed N_F to investigate the performance when N_F is reduced.

(c) We changed N_V to investigate the performance when N_V is increased.

Table 5 presents the quantitative results for the ablation studies. The GMGS decreased by 0.283 points under condition (a). This result indicates that the introduction of the SFM to handle both image and physical features was beneficial to the performance. Under conditions (b) and (c), the scores of each metric fluctuated slightly. We selected condition (d) as the proposed method because the

Table 5. Quantitative results for ablation studies. (a) w/o $F_{t-k+1:t}$ (b) $(N_V, N_F) = (1, 1)$ (c) $(N_V, N_F) = (2, 2)$ (d) $(N_V, N_F) = (1, 2)$.

Conditions	GMGS↑	TSS$_{\geq\mathrm{M}}$↑	BSS$_{\geq\mathrm{M}}$↑
(a)	0.220±0.116	0.198±0.371	-1.77±0.225
(b)	0.516±0.089	0.485±0.082	0.052±1.05
(c)	0.563±0.070	0.551±0.123	0.011±0.965
(d)	0.503±0.059	0.530±0.112	0.082±0.974

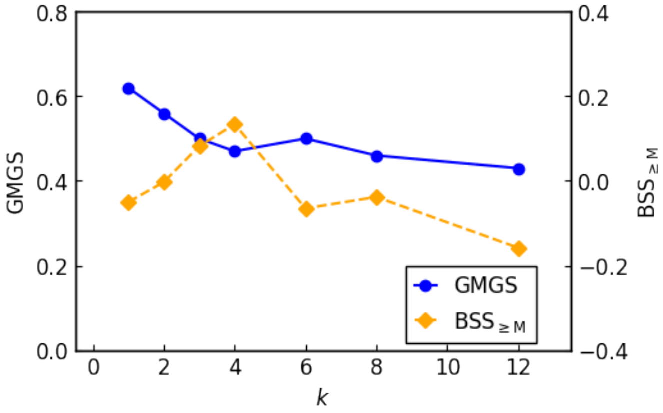

Fig. 5. GMGS and BSS$_{\geq\mathrm{M}}$ plotted against k.

scores were well balanced. The above results indicate that handling both magnetograms and physical features as input was beneficial to performance.

To investigate the effect of past images on prediction performance, we evaluated GMGS and BSS$_{\geq\mathrm{M}}$ for various k. For example, $k = 4$ means that $\boldsymbol{x}_{t-3:t}$ was used as the model input. The results in Fig. 5 shows that the highest values of GMGS and BSS$_{\geq\mathrm{M}}$ occur at $k = 1$ and $k = 4$, respectively. This indicates that $k = 1$ is sufficient to maximize GMGS. However, it also indicates that we need to consider an appropriate k in order to balance it with BSS$_{\geq\mathrm{M}}$.

5.5 Error Analysis

Table 6 presents the confusion matrix for our method using 2017 test set. For the X-class, there were 20, 77, 16, and 7878 true positive (TP), false positive (FP), false negative (FN), and true negative (TN) samples, respectively.

Table 7 categorizes the failed cases. We define the influence of failure cases on GMGS as follows:

$$\mathrm{GMGS}_{\mathrm{Influence}} = \frac{c_{ij}(s_{ii} - s_{ij})}{N}, \tag{29}$$

where c_{ij} and s_{ij} denote the element (i, j) for the confusion matrix and for the GMGS score matrix [6], respectively. Table 7 indicates that the bottleneck is

Table 6. Confusion matrix for 2017 test set.

		Predicted Flare Class			
		O	C	M	X
Observed Flare Class	O	7269	210	34	12
	C	84	150	29	22
	M	18	50	34	43
	X	1	0	15	20

Table 7. The error analysis by $\mathrm{GMGS}_{\mathrm{Influence}}$ for 2017 test set.

Observed Class	Predicted Class	$\mathrm{GMGS}_{\mathrm{Influence}}$
X	M	0.1335
M	C	0.0885
C	O	0.0578
M	O	0.0442
X	O	0.0114

the misprediction of X-class for M-class flares. This suggests that methods to alleviate this bottleneck will effectively enhance the prediction performance.

6 Conclusion

In this paper, we proposed the Flare Transformer (FT), a method for predicting the maximum solar flare class that will occur within 24 h. The following contributions of this study can be emphasized:

- We proposed the FT method, which handles both line-of-sight magnetograms and physical features through the MM and SFM.
- We introduced transformer attention mechanism [27] to model temporal relationships between input features.
- We introduced the GMGS and BSS losses to balance the two major metrics in solar flare prediction.
- We have demonstrated that our model gives better predictions than can be achieved by human experts in terms of GMGS and TSS.

Acknowledgements This work was partially supported by JSPS KAKENHI Grant Number 20H04269 and NEDO.

References

1. Angryk, R.A., Martens, P.C., Aydin, B., Kempton, D., Mahajan, S., Basodi, S., Ahmadzadeh, A., Cai, X., Boubrahimi, F., Hamdi, M.: Multivariate Time Series Dataset for Space Weather Data Analytics. Scientific data **7**(1), 1–13 (2020)
2. Bhattacharjee, S., Alshehhi, R., Dhuri, D., et al.: Supervised Convolutional Neural Networks for Classification of Flaring and Nonflaring Active Regions using Line-of-sight Magnetograms. The Astrophysical Journal **898**(2), 98 (12pp) (2020)
3. Box, G., Jenkins, G., Reinsel, G., Ljung, G.: Time Series Analysis: Forecasting and Control. John Wiley & Sons (2015)
4. Cinto, T., Gradvohl, S., Coelho, P., Silva, A.: A Framework for Designing and Evaluating Solar Flare Forecasting Systems. Monthly Notices of the Royal Astronomical Society **495**(3), 3332–3349 (2020)
5. Esling, P., Agon, C.: Time-series data mining. ACM Computing Surveys **45**(1), 1–34 (2012)
6. Gandin, L., Murphy, A.: Equitable Skill Scores for Categorical Forecasts. Monthly Weather Review **120**(2), 361–370 (1992)
7. Georgoulis, M., Bloomfield, S., et al.: The flare likelihood and region eruption forecasting (FLARECAST) project: flare forecasting in the big data & machine learning era. Journal of Space Weather and Space Climate **11**, A39 (37pp) (2021)
8. Huang, X., Wang, H., Xu, L., Liu, J., Li, R., Dai, X.: Deep Learning Based Solar Flare Forecasting Model. I. Results for Line-of-sight Magnetograms. The Astrophysical Journal **856**(1), 7 (11pp) (2018)
9. Hyndman, R., Koehler, A., et al.: Forecasting with Exponential Smoothing: The State Space Approach. Springer Science & Business Media (2008)
10. Kubo, Y., Den, M., Ishii, M.: Verification of Operational Solar Flare Forecast: Case of Regional Warning Center Japan. Journal of Space Weather and Space Climate **7**, A20 (16pp) (2017)
11. Kusano, K., Iju, T., Bamba, Y., Inoue, S.: A Physics-based Method that can Predict Imminent Large Solar Flares. Science **369**(6503), 587–591 (2020)
12. Lemen, J., Akin, D., Boerner, P., Chou, C., Drake, F., Duncan, W., Edwards, G., et al.: The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). In: The Solar Dynamics Observatory, pp. 17–40. Springer (2011)
13. Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.X., Yan, X.: Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. In: NeurIPS. vol. 32, pp. 5243–5253 (2019)
14. Murray, S., Bingham, S., Sharpe, M., Jackson, D.: Flare forecasting at the met office space weather operations centre. Space Weather **15**(4), 577–588 (2017)
15. Nishizuka, N., Kubo, Y., Sugiura, K., Den, M., Ishii, M.: Reliable Probability Forecast of Solar Flares: Deep Flare Net-Reliable (DeFN-R). The Astrophysical Journal **899**(2), 150 (8pp) (2020)
16. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., et al.: Solar Flare Prediction Model with Three Machine-learning Algorithms using Ultraviolet Brightening and Vector Magnetograms. The Astrophysical Journal **835**(2), 156 (10pp) (2017)
17. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Ishii, M.: Deep flare net (DeFN) model for solar flare prediction. The Astrophysical Journal **858**(2), 113 (8pp) (2018)
18. Ong, B.T., Sugiura, K., Zettsu, K.: Dynamically Pre-trained Deep Recurrent Neural Networks using Environmental Monitoring Data for Predicting PM2.5. Neural Computing and Applications **27**(6), 1553–1566 (2016)

19. Park, E., Moon, Yong-Jae, S., Yi, K., Lim, D., et al.: Application of the Deep Convolutional Neural Network to the Forecast of Solar Flare Occurrence Using Full-disk Solar Magnetograms. The Astrophysical Journal **869**(2), 91 (6pp) (2018)
20. Pesnell, W., Thompson, B., Chamberlin, P.: The Solar Dynamics Observatory (SDO). In: The Solar Dynamics Observatory, pp. 3–15. Springer (2011)
21. Rangapuram, S., Seeger, M., Gasthaus, J., Stella, L., et al.: Deep State Space Models for Time Series Forecasting. In: NeurIPS. vol. 31, pp. 7785–7794 (2018)
22. Re, S.: Solar storm; how to calculate insured / reinsured losses? Space Weather Workshop (2016)
23. Salinas, D., Flunkert, V., Gasthaus, J., Januschowski, T.: DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting **36**(3), 1181–1191 (2020)
24. Scherrer, P., Schou, J., Bush, R., Kosovichev, A., Bogart, R., Hoeksema, J., Liu, Y., Duvall, T., et al.: The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO). Solar Physics **275**(1), 207–227 (2012)
25. Tang, R., Liao, W., Chen, Z., Zeng, X., Wang, J.s., Luo, B., Chen, Y., Cui, Y., et al.: Solar Flare Prediction Based on the Fusion of Multiple Deep-learning Models. The Astrophysical Journal Supplement Series **257**(2), 50 (13pp) (2021)
26. Tashman, L.: Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting **16**(4), 437–450 (2000)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., et al.: Attention is all you need. In: NeurIPS. vol. 30, pp. 5998–6008 (2017)
28. Wen, R., Torkkola, K., Narayanaswamy, B., Madeka, D.: A Multi-horizon Quantile Recurrent Forecaster. arXiv preprint arXiv:1711.11053 (2017)
29. Wu, S., Xiao, X., Ding, Q., Zhao, P., et al.: Adversarial Sparse Transformer for Time Series Forecasting. In: NeurIPS. vol. 33, pp. 17105–17115 (2020)
30. Zhou, H., Zhang, S., Peng, J., Zhang, S., et al.: Informer: Beyond Efficient Transformer for Long Sequence Time-series Forecasting. In: AAAI. pp. 11106–11115 (2021)