Transformer Model for Multivariate Time Series Classification: A Case Study of Solar Flare Prediction

Khaznah Alshammari[1,2][0009−0005−4435−9642], Shah Muhammad Hamdi[3][0000−0002−9303−7835], and Soukaina Filali Boubrahimi[3][0000−0001−5693−6383]

[1] New Mexico State University,Las Cruces, NM 88001, USA
[2] Northern Border University, Rafha, Saudi Arabia
kalshamm@nmsu.edu
[3] Utah State University, Logan, UT 84322, USA
{s.hamdi,soukaina.boubrahimi}@usu.edu

Abstract. Classifying solar flares is essential for understanding their impact on space weather forecasting. We propose a novel approach using a multi-head attention and transformer mechanism to classify multivariate time series (MVTS) instances of photospheric magnetic field parameters of the flaring events in the solar active regions. Attention mechanisms and transformer architectures capture complex temporal dependencies and interactions among features in multivariate time series data. Our model simultaneously attends to relevant features and learns their dependencies, enabling accurate classification of solar flare events. We evaluated our approach on *SWAN-SF*, the largest MVTS dataset for predicting solar flares, and compared its performance against several state-of-the-art methods. The experimental results demonstrate that our approach achieves superior classification performance, even when dealing with a highly imbalanced dataset characterized by the scarcity of major flaring events. These findings highlight the effectiveness of attention mechanisms and transformer models in learning discriminatory features from MVTS-based space weather data.

Keywords: Solar flares · Multivariate time series · Attention-based framework · Classification · Space weather forecasting.

1 Introduction

A solar flare is an intense, localized eruption of electromagnetic radiation in the Sun's atmosphere. Flares occur in active regions and are often accompanied by coronal mass ejections, solar particle events, and other solar phenomena. The occurrence of solar flares varies with the 11-year solar cycle. Solar flares tend to be more frequent and intense during periods of high solar activity, which coincide with the solar maximum phase of the solar cycle. Solar flares result from the abrupt release of accumulated magnetic energy within the Sun's atmosphere. This energy can be stored in twisted magnetic fields above sunspots or

in other active regions. When magnetic energy is released, it heats the surrounding plasma to millions of degrees Celsius and accelerates particles to near the speed of light. Solar flares can produce a wide range of electromagnetic radiation, from radio waves to X-rays and gamma rays. Solar flares can have a significant impact on Earth and the space environment. They can cause radio blackouts, damage power grids, and disrupt satellite communications. Solar flares can also produce high-energy particles that can pose a hazard to astronauts and aircraft crews [14,18,19]. Effective predictions of solar flares are facilitated by employing time series modeling on magnetic field data collected by The Solar Dynamics Observatory's Helioseismic Magnetic Imager (HMI). Consequently, spatiotemporal magnetic field data is mapped into multiple instances of Multivariate Time Series (MVTS) [7]. Each MVTS instance contains solar magnetic field parameters such as flux, current, helicity, and Lorentz force. The time series associated with these parameters are derived from two distinct time windows: the observation window, which encompasses the recording of magnetic field parameter values, and the subsequent prediction window, corresponding to the period when peak X-ray flux was observed. The instances are then labeled into six classes: Q, A, B, C, M, and X. "Q" represents flare quiet active regions, while the other labels represent flaring events with increasing intensity. Notably, X and M-class flares denote the most intense flaring events. Recent advances in Multivariate Time Series (MVTS) models have demonstrated their superior effectiveness in predicting solar flaring activities when compared to earlier models that relied on single timestamps for magnetic field vector classification [7]. MVTS-based models for flare prediction can be broadly categorized into two main groups. The first category is the statistical feature-based method [16]. In this approach, low-dimensional representations of MVTS instances are computed by aggregating summary statistics from the individual univariate time series components. Subsequently, traditional classifiers such as k-nearest Neighbors (kNN) and Support Vector Machines (SVM) are trained using these labeled MVTS representations. The second category comprises end-to-end deep learning-based methods [24], which utilize Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) based deep sequence models. These models learn by sequentially inputting vectors representing magnetic field parameters into the cells of the sequence model. The cell weights are optimized through backpropagation based on gradient descent. However, a limitation of these models is that they can only leverage the temporal dimension of the MVTS instances, resulting in suboptimal classification performance due to their limited ability to exploit the underlying patterns within the data. Vaswani et al. proposed the Transformer model, a neural network architecture based solely on self-attention mechanisms, to address the limitations of previous models [29]. The introduction of the transformer model marked a significant breakthrough in the field of natural language processing (NLP) and served as the cornerstone for numerous subsequent advancements, including cutting-edge language models such as BERT [13] and GPT [32]. One of its primary advantages lies in its efficiency in capturing long-range dependencies in data, all while allowing for parallel processing. This leads to faster training

and inference times in comparison to previous models. The effectiveness of the transformer model, makes it a powerful choice for MVTS classification, leveraging its ability to capture long-range dependencies and handle multi-variable, temporal data effectively [29]. In our study, we aim to explore an alternative approach using attention and transformer techniques. By harnessing the power of self-attention mechanisms in transformers, our objective is to capture the extended temporal relationships among magnetic field parameters within the MVTS data, improving solar flare classification performance and deepening our understanding of these events. In this paper, we propose an attention-based model for the MVTS classification by leveraging the self-attention mechanisms to improve the MVTS classification performance. Our experimental results of our model demonstrate a test score of 70% on the solar flare MVTS dataset when using the proposed attention-based model, outperforming the baselines by more than 10%.

2 Related Work

Historically, systems for predicting solar flares heavily relied on human expertise and manual inputs. One early system known as THEO, implemented by the Space Environment Center (SEC) of NOAA back in 1987, required human intervention to input sunspot characteristics to categorize flare classes [22]. However, as recent NASA missions have generated a wealth of magnetic field data, the focus has shifted towards data-driven approaches, moving away from purely theoretical models. These data-driven approaches can be broadly categorized into linear and nonlinear statistical models, depending on the nature of the dataset used. These models can further be divided into line-of-sight magnetogram-based and full-disk photospheric vector magnetogram-based models. Line-of-sight magnetogram data captures only the component of the magnetic field along the line of sight, while full-disk photospheric vector magnetic field data provides a more comprehensive magnetic field state of solar active regions. Linear statistical models aimed to identify highly correlated magnetic field features associated with flare occurrences. For instance, Cui et al. [11] used line-of-sight magnetogram data to establish correlation-based statistical relationships between magnetic field parameters and flare events. Even before the launch of the Solar Dynamics Observatory (SDO), Leka et al. [20] utilized linear discriminant analysis (LDA) to classify flaring events using vector magnetogram data from the Mees Solar Observatory. In contrast, nonlinear statistical models employed a range of machine learning classifiers such as logistic regression, decision trees, neural networks, support vector machines (SVM), and more. For example, Song et al. [28] and Yu et al. [31] applied various classifiers to line-of-sight magnetogram-based datasets. Bobra et al. [9] utilized SVM with SDO-based vector magnetogram data for flare classification, while Nishizuka et al. [26] compared the performance of k-Nearest Neighbors (kNN), SVM, and Extremely Randomized Tree (ERT) on both line-of-sight and vector magnetogram data. Furthermore, Convolutional Neural Networks (ConvNets) have been employed for solar flare prediction using

SDO AIA/HMI images [21, 33]. Recently, Angryk et al. introduced a novel approach to solar flare prediction based on temporal windows, which represents an extension beyond the earlier single timestamp-based models [7]. In their method, they employed a Multivariate Time Series (MVTS) dataset consisting of active regions, which recorded magnetic field data over a predefined observation period at a consistent sampling rate. Each instance in this dataset was labeled based on the flare classes that occurred after a specified prediction time. Other efforts, such as Hamdi et al. [17] and Muzaheed et al. [24], utilized statistical summarization, decision trees, and Long Short-Term Memory (LSTM)-based deep sequence modeling for flare prediction. Furthermore, Alshammari et al. [6] addressed the task of forecasting future values of magnetic field parameters within the MVTS representations. This involved predicting forthcoming values based on past data in the MVTS dataset. The transformer model, introduced by Vaswani et al. [29], offers several strengths for MVTS classification, including long-range dependency modeling. The transformer model can capture long-range dependencies in the data, which is effective for MVTS classification. Parallel computation of the transformer model makes it efficient for training and inference on large MVTS datasets. Transformer models can support contextual learning, enabling them to discern contextual relationships between magnetic field parameters without the need for explicit sequential processing. This is important for the classification of MVTS instances, as the context of a particular time step can be informative for predicting the occurrence of a solar flare.

3 Methodology

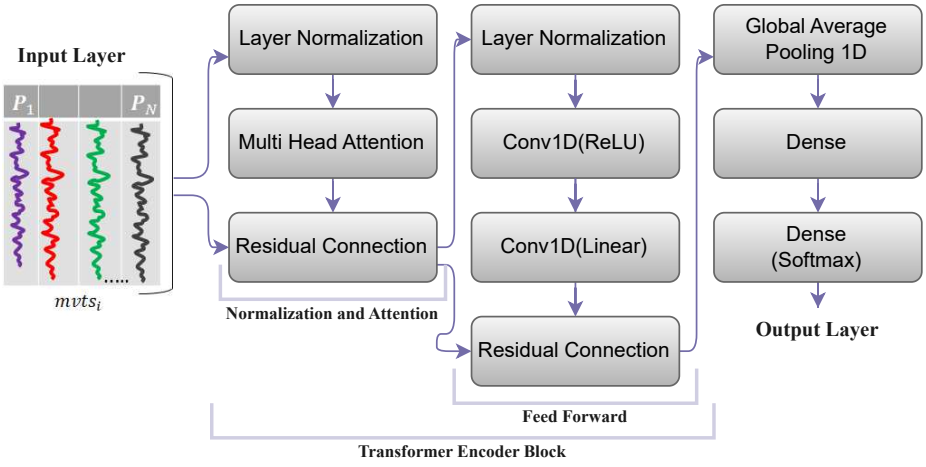

Fig. 1: Transformer Model for MVTS Classification

3.1 Notations

The solar event instance i is represented by an MVTS instance $mvts_i$. The MVTS instance $mvts_i \in \mathbb{R}^{T \times N}$ is a collection of individual time series of N magnetic field parameters, where each time series contains periodic observation values of the corresponding parameter for an observation period T. The MVTS instance can be expanded as $mvts_i = \{v_{t_1}, v_{t_2}, ., ., ., v_{t_T}\}$, where $v_{t_i} \in \mathbb{R}^N$ represents a timestamp vector.

3.2 Data Preprocessing and Normalization

The magnetic field parameter values are recorded in different scales, so we perform z-score normalization of each individual time series of each MVTS instance. At $mvts_i$, parameter-based individual time series are denoted by $P_1, P_2, \ldots, P_N$. For each individual time series P_j, we perform z-normalization as follows.

$$x_k^{(j)} = \frac{x_k^{(j)} - \mu^{(j)}}{\sigma^{(j)}}$$

Here, $x_k^{(j)}$ is the k-th value of the time series P_j, where $1 \leq k \leq T$, $\mu^{(j)}$ is the mean of time series P_j, and $\sigma^{(j)}$ is the standard deviation of the time series P_j. We apply the z-normalization for each partition individually. When partition i is used for training and partition j is used for test, we perform above z-normalization independently in the MVTS instances of partition i and j.

3.3 Transformer Model for MVTS Classification

In this work, we harness an attention-based model (transformer) to enhance the classification performance in an MVTS-based solar flare dataset. Within our model, we have designed the transformer encoder block. The foundation for this approach is rooted in the work of Vaswani et al. [29], where they introduced the transformer. The transformer architecture comprises both an encoder and a decoder, each comprising multiple layers that integrate self-attention and feed-forward neural networks. In our specific application, we primarily focus on the encoder component. This encoder is responsible for processing the input sequence, which, in the context of our study, corresponds to the solar flare data. The encoder structure consists of a stack of identical layers, with each layer housing two sub-layers:

- Self-attention layer: This layer plays a pivotal role by enabling each timestamp within the input time series to attentively consider all other timestamps within the same sequence. This mechanism empowers the layer to capture intricate temporal dependencies between individual timestamps and generate context-aware representations for each timestamp.

– Feed-forward neural network layer: Following the self-attention mechanism, a feed-forward neural network layer is independently applied to each timestamp representation. This layer introduces non-linearity into the model, allowing it to incorporate additional information and enhance its overall performance.

In this model, we incorporate the transformer encoder block and leverage the advantages of the multi-head attention architecture, a critical component of the transformer model. This architecture empowers the model to simultaneously focus on various segments of the input sequence, thereby enhancing its capacity to capture intricate temporal dependencies and extract pertinent features. By employing multiple attention heads, our model can acquire diverse representations and attend to distinct aspects of the input data concurrently. In the context of classifying MVTS data, multi-head attention offers several significant benefits:

– Enhanced Representational Capacity: Multi-head attention permits the model to attend to different portions of the input sequence in parallel, facilitating the capture of both local and global dependencies effectively. This capability empowers the model to discern complex patterns within the time series data, ultimately leading to improved classification performance.
– Robustness to Variable-Length Sequences: MVTS data frequently comprises sequences of varying lengths. Multi-head attention adeptly manages sequences of different lengths by assigning varying attention weights to different segments of the input. This adaptability enables the model to accommodate sequences with differing lengths without compromising its classification accuracy.

The key equations governing the multi-head attention mechanism are presented and explained in [29]. Our model, illustrated in Figure 1 is described in algorithms 1 and 2. Algorithm 1 operates as follows:

Algorithm 1 MVTS Transformer Encoder

function TRANSFORMER_ENCODER($inputs$, $head\_size$, num_heads, ff_dim)
  $x \leftarrow \text{LAYERNORMALIZATION}(inputs, \epsilon = 1e-6)$
  $x \leftarrow \text{MULTIHEADATTENTION}(x, x, key\_dim = head\_size, num\_heads = num\_heads)$
  $res \leftarrow x + inputs$
  $x \leftarrow \text{LAYERNORMALIZATION}(res, \epsilon = 1e-6)$
  $x \leftarrow \text{CONV1D}(x, filters = ff\_dim, kernel\_size = 1, activation = \text{"}relu\text{"})$
  $x \leftarrow \text{CONV1D}(x, filters = inputs.shape[-1], kernel\_size = 1)$
  return $x + res$
end function

1. Layer Normalization: The tensor representation of MVTS instances is first normalized along each feature dimension by passing it through a layer normalization layer.

2. Self-Attention: The normalized tensor is then fed into a multi-head attention layer, where a self-attention mechanism is applied. Each attention head attends to different parts of the input sequence and learns to capture distinct relationships between time steps. The output of the attention layer retains the same shape as the input.
3. Residual Connection: The output of the multi-head attention layer is element-wise added to the original input tensor (inputs). This residual connection facilitates the direct flow of gradients from the input to the output, easing the learning process for the model.
4. Feed-forward layer: The result of the residual connection is passed through another layer normalization layer.
5. Convolutional Layer: A 1D convolutional layer with ff_dim filters and kernel size 1 is applied to the normalized tensor. This layer acts as a feed-forward neural network layer, applying non-linear transformations independently to each position in the sequence.
6. Second Convolutional Layer: Another 1D convolutional layer with inputs of shape[-1] filters and kernel size 1 is applied to the result obtained from the previous layer.
7. Residual Connection: The output of the second convolutional layer is element-wise added to the result obtained from the first residual connection layer.
8. Final Output: The sum of the previous residual connection and the original input tensor (inputs) is returned as the final output.

Algorithm 2 Build MVTS Transformer(Attention) Model

function BUILD_TRANSFORMER_MODEL($input\_shape$, $head\_size, num\_heads, ff\_dim$, $num\_transformer\_blocks$, mlp_units)
  $n\_classes \leftarrow$ LENGTH($unique\_y\_train$)
  $inputs \leftarrow$ INPUT($shape = input\_shape$)
  $x \leftarrow inputs$
  for $i \leftarrow 1$ **to** $num\_transformer\_blocks$ **do**
    $x \leftarrow$ TRANSFORMER_ENCODER(x, $head\_size$, num_heads, ff_dim)
  end for
  $x \leftarrow$ GLOBALAVERAGEPOOLING1D(x, $data\_format = "channels\_first"$)
  for dim **in** mlp_units **do**
    $x \leftarrow$ DENSE($x, dim, activation = "relu"$)
  end for
  $outputs \leftarrow$ DENSE($x, n\_classes, activation = "softmax"$)
  return MODEL($inputs, outputs$)
end function

Algorithm 2 incorporates several parameters, each described as follows: $input\_shape$ specifies the shape of the input data, $head\_size$ determines the size of each attention head in the transformer, num_heads denotes the number of attention heads in the transformer, ff_dim represents the dimension of the feed-forward

network in the transformer, $num\_transformer\_blocks$ indicates the number of transformer blocks to be stacked, and mlp_units is a list of integers specifying the number of units in each MLP layer. Within the algorithm, it first determines the number of classes ($n\_classes$) based on the unique labels present in the training data. It then defines the input layer and sets it as the current layer, denoted as x. The algorithm proceeds by applying the transformer encoder block through the $transformer\_encoder$ function. After the transformer encoder blocks, a global average pooling layer is applied to reduce the spatial dimensions of the data. Subsequently, a series of MLP layers are implemented as specified by the mlp_units parameter, with each layer employing $ReLU$ activation. Finally, an output layer is added with $n\_classes$ units and a $softmax$ activation function for classification.

4 Experiments

In this section, we present our experimental findings, where we compare the performance of our model with five other MVTS-based flare prediction baselines using a benchmark dataset. We implemented our attention-based MVTS classifier using TensorFlow on the A100 Nvidia GPU. The hyper-parameters were found by random hyper-parameter search, and set as $head\_size$=256, num_heads=4, ff_dim=4, $num\_transformer\_blocks$=10, mlp_units= 64. The source code of our model, along with the experimental dataset, is available in our GitHub repository.[4]

4.1 Evaluation Metrics

We used True Skill Statistic (TSS) as a performance measure for our experiments. The True Skill Statistic (TSS) takes into account both the hits and false alarms in the prediction. It is calculated as

$$TSS = \frac{TP}{TP + FN} - \frac{FP}{FP + TN}$$

where TP is the number of true positives (correct predictions of flares), FN is the number of false negatives (missed predictions of flares), FP is the number of false positives (incorrect predictions of flares), and TN is the number of true negatives (correctly predicted non-flares). The TSS ranges from -1 to 1, where a value of 1 represents a perfect prediction, a value of 0 represents a random prediction, and -1 indicates that the model is wrong in all of its predictions [8]. We use TSS as a performance metric because it has been used frequently to report the performance of machine learning models for the prediction of rare events, e.g., solar flares [1,7,17]. TSS can accurately measure the model's ability to distinguish between the classes, regardless of how common or rare they are. TSS is widely used in machine learning and statistical modeling, especially for tasks such as

[4] `https://github.com/Kalshammari/Transformer-Model.git`

binary classification [15]. One advantage of using TSS as a performance metric in datasets with high-class imbalance is that it takes into account both the true positive rate and the true negative rate of the classifier, which is important when the classes are imbalanced [15]. TSS can also be used to compare models with different thresholds for presence-absence predictions [3].

4.2 SWAN-SF Benchmark Data Set

As the benchmark dataset of our experiments, we used the MVTS-based solar flare prediction dataset SWAN-SF published in [7]. The SWAN-SF benchmark dataset is a collection of multivariate time series (MVTS) data instances that facilitate unbiased flare forecasting. The MVTS instances of the SWAN-SF benchmark dataset are labeled by five different flare classes, namely, GOES X, M, C, and B, and a non-flaring class denoted by Q. Class Q includes flare-quiet events and GOES A-class events. The dataset comprises five temporally segmented partitions and is designed in a way that each partition includes approximately the same number of X- and M-class flares. Table 1 shows the partition-wise label statistics of the SWAN-SF dataset. The dataset contains various time series parameters derived from solar photospheric magnetograms along with NOAA's flare history of active regions. The magnetograms and their metadata are obtained from the Spaceweather HMI Active Region Patch (SHARP) data product. The magnetic field parameters are physics-based and were recalculated and enhanced for validation purposes. Each MVTS instance in the dataset is made up of a 24-time series of active region magnetic field parameters (the full list can be found in [4, 9]. The time series instances are recorded at 12-minute intervals for a total duration of 12 hours (that results in 60-time steps). In this paper, $T = 60$ is used to denote the number of observation time steps, and $N = 24$ to denote the number of magnetic field parameters. In this study we use all 24 magnetic field parameters. In our experiments of feature selection from MVTS data, we conduct the binary classification between flaring and non-flaring AR, where we consider flaring AR (class X and M) to be in a positive class and non-flaring Active Regions (class Q) to be in the negative class. We removed the B and C class events since the opposite event classes (X + M vs Q) help in contrastive learning. The removal of B and C class flares for maximizing flare prediction performance was also suggested by the experimental findings of multiple previous studies [2, 5, 9, 17, 27].

4.3 Baseline Models

We evaluated our model with five other baselines.

1. **Long Short-Term Memory (LSTM)** The LSTM-based approach was proposed by Muzaheed et. al. [24]. Each MVTS instance was considered as a T -length sequence of $x^{<t>} \in \mathbb{R}^N$ timestamp vectors. After sequentially feeding the LSTM model with each timestamp vector, the last hidden representation was considered as the MVTS representation. As suggested by the

paper, we set the number of cell state and hidden state dimensions to 64, the number of training epochs to 500, and the learning rate in stochastic gradient descent to 0.01.

2. **Support Vector Machine (SVM)** SVM is known for its ability to handle linear and non-linear data effectively, making it a versatile choice for various applications. It employs support vectors, which are data points closest to the decision boundary, to determine the orientation and placement of the hyperplane. This approach allows SVM to excel in complex and high-dimensional datasets [10].
3. **Canonical Interval Forest (CIF)** The time series forest (TSF) classifier, known for its high performance, quick training, and prediction, is commonly regarded as a powerful interval method proposed by [23].
4. **Multiple Representations SEQuence Learner (MRSEQL)** MRSEQL, proposed by [25], is a robust univariate time series classifier that trains on features derived from multiple symbolic representations of time series. These representations include Symbol Aggregation Approximation (SAX) and Symbol Fourier Approximation (SFA), which are used with linear classification models (logistic regression).
5. **MINImally RandOm Convolutional KErnels Transform (MINIROCKET)** MINIROCKET is a fast and accurate algorithm for time series classification. It is a (nearly) deterministic reformulation of the ROCKET algorithm, which is a state-of-the-art algorithm for time series classification [12].

4.4 Train/validation/test splitting method

The SWAN-SF dataset has a temporal coherence property that measures how stable and consistent the magnetic field structures of a solar active region are over time. It poses a challenge for predicting rare events such as solar flares using time series data. It requires that the predictions for a given time point are in agreement with past and future predictions. If temporal coherence is ignored, the model performance may be artificially inflated. This problem stems from the data collection method and affects the data splitting into training, validation, and testing sets. To address the issue of temporal coherence, we use different time-segmented partitions of the dataset for training and testing samples. This is the reason why the SWAN-SF dataset has multiple non-overlapping partitions. Table 1 shows each partition statistics. By using different partitions of SWAN-SF for training and testing, we avoid testing the model on time series that are partly identical to those used for training [1]. In this study, we use the following settings: partition 1 for training and validation and partition 2 for testing, partition 2 for training and validation and partition 3 for testing, partition 3 for training and validation and partition 4 for testing, partition 4 for training and validation and partition 5 for testing, and partition 5 for training and validation and partition 1 for testing.

Table 1: Event type statistics of each partition of the SWAN-SF dataset

Flare Type	**Partitions**				
	P1	P2	P3	P4	P5
Q	60,130	73,368	34,762	43,294	62,688
B	5,692	4,978	685	846	5,924
C	6,416	8,810	5,639	5,956	5,763
M	1,089	1,329	1,288	1,012	971
X	165	72	136	153	19
sum	**73,492**	**88,557**	**42,510**	**51,261**	**75,365**

4.5 Binary classification performance

Binary classification plays a significant role in distinguishing major flaring events from minor flaring events or flare quiet events. In this experiment, we focus on classifying X and M class MVTS instances as flaring events, while considering all other instances (Q) as non-flaring events. Figure 2 depicts the binary classification performances of all models. The results demonstrate that the transformer-based MVTS model outperforms all other baseline models, and achieves an average improvement of approximately 8% to 20% compared to the second-best performing MINIROCKET algorithm. These findings highlight the superior performance of our model in binary classification. This consistency reinforces the efficacy and reliability of our Transformer-based model in accurately predicting flaring events.

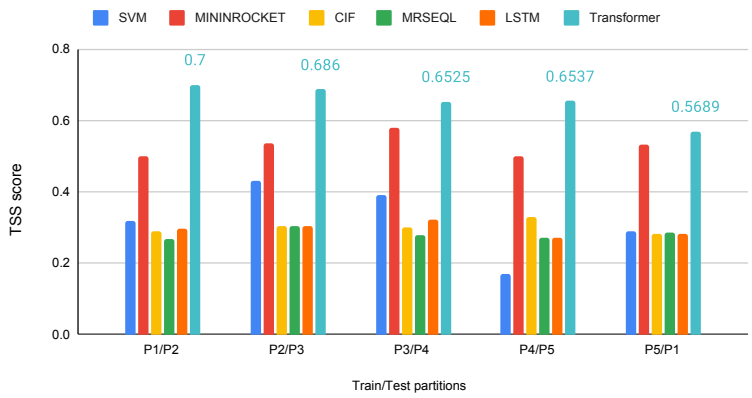

Fig. 2: Binary classification performance of all baselines compared to the transformer model.

4.6 Ablation Study of the Transformer-base MVTS Classification Model

To get a better understanding of the effectiveness of the different layers in our model, we conducted several experiments to evaluate the significance of various layers. First, we evaluated the importance of the self-attention mechanism by removing it from the model architecture and comparing the results. The removal of the attention mechanism (when partition 1 was used for training and partition 2 for testing) led to a noticeable drop in the TSS score, from 70% to 54%. This outcome highlights the significant role played by the multi-head attention layer in capturing relevant patterns and relationships within the MVTS data. Second, we examined the impact of layer normalization by removing it from the model. This resulted in a decrease in TSS from 70% to 44%. This finding underscores the importance of layer normalization in maintaining the model's performance and stability. Finally, we investigated the effect of the 1D convolutional layers. When these layers were removed from the model, there was a significant drop in TSS from 70% to 51%. This result demonstrates the crucial role played by the 1D convolutional layers in capturing important temporal features and contributing to the overall performance of the model. The ablation study provided valuable insights into the contributions of different layers in our model. The significant decrease in TSS upon removing the attention mechanism, layer normalization, and 1D convolutional layers highlights their importance in capturing relevant patterns, maintaining stability, and extracting essential temporal features. These findings underscore the effectiveness and significance of each layer in our model architecture.

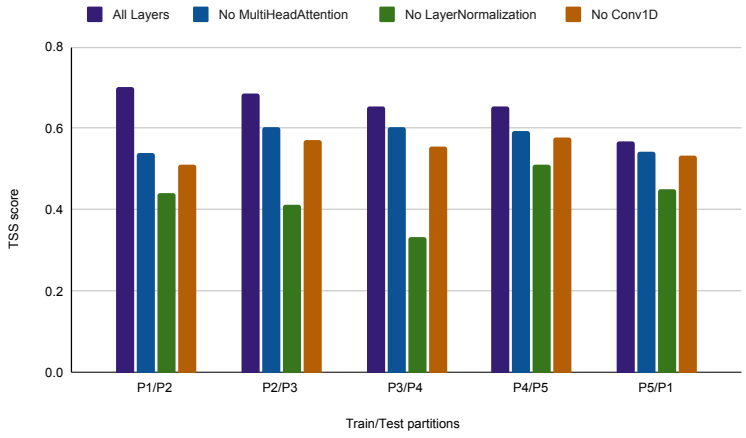

Fig. 3: Ablation Study: The Contributions of Model Components in MVTS Classification of Solar Flares.

Table 2: Experimental Results (TSS scores) for Different Hyperparameters Values on (Train/Test) Partitions.

Head Size	Num Heads	FF Dim	Num of Transformer Blocks	MLP Units	TSS (P1/P2)	TSS (P2/P3)	TSS (P4/P5)	TSS (P5/P1)
256	4	4	10	64	**0.70**	**0.69**	**0.58**	0.57
512	4	4	10	64	0.62	0.68	0.52	0.63
512	8	8	20	128	0.68	0.68	0.57	**0.64**
128	2	2	5	32	0.69	0.65	0.47	0.29
256	2	2	5	32	0.58	0.65	0.41	0.45
256	4	4	10	128	0.58	0.60	0.32	0.46

4.7 The Impact of Different Hyperparameters Values on The Model Performance

Table 2 presents the performance of various Transformer model configurations, highlighting key hyperparameters such as head size, number of heads, feed-forward dimensions, number of Transformer blocks, and MLP units. The Total Sum of Squares (TSS) scores across different train/test partitions (P1/P2, P2/P3, P4/P5, and P5/P1) demonstrate the model's effectiveness in capturing data variance. For the first row in the table, the model configuration includes an attention head size of 256, 4 attention heads, a feed-forward dimension of 4, and 10 Transformer blocks. Additionally, the MLP (Multi-Layer Perceptron) units are set to 64. The computational complexity for this configuration is approximately $O(10 \cdot (n^2 \cdot 256 + n \cdot 256^2))$, where n represents the sequence length. This complexity estimate indicates how the computational cost grows with the sequence length (n) and model size (256), helping to understand the resource requirements for this specific model setup.

5 Discussion

We acknowledge that the application of the vanilla transformer architecture is not novel in a methodological sense, we believe that the contribution of our study lies in its specific adaptation to the solar flare prediction domain. The utilization of self-attention mechanisms within the transformer framework, tailored to the characteristics of solar flare MVTS datasets, addresses unique challenges in time series classification. Our primary focus was to explore the effectiveness of self-attention mechanisms in capturing long-range dependencies and intricate patterns inherent in solar flare data. We believe that the context-specific adaptation of the transformer architecture contributes valuable insights to the solar flare prediction community. The impact of this study in the field of space weather forecasting is significant. Accurate prediction of solar flares is important for mitigating the adverse effects of space weather on satellite communications, power grids, and other critical infrastructure. By leveraging the advanced capabilities of Transformer models, this research provides a robust framework for enhancing the prediction accuracy of solar flare events. The use of self-attention

mechanisms enables the model to capture intricate temporal dependencies and interactions among multiple magnetic field parameters, which are essential for understanding the complex dynamics of solar flares. The proposed model's ability to handle large-scale multivariate time series data and its applicability to real-world scenarios make it a practical tool for operational space weather forecasting. By addressing the limitations of previous models and demonstrating superior performance, this research contributes to the development of more reliable and effective space weather prediction systems. The Transformer model is important, particularly in the context of solar flare prediction. The self-attention mechanism used in the Transformer model allows it to focus on different parts of the input sequence, providing insights into which features and time steps are most influential in making predictions. This capability can help identify key magnetic field parameters and their interactions that contribute to the occurrence of solar flares. Furthermore, by analyzing the attention weights, researchers can gain a better understanding of the physical mechanisms underlying solar flare events. The model's ability to capture long-range dependencies and complex temporal relationships in multivariate time series data makes it a powerful tool for studying the dynamics of solar active regions. This can lead to improved predictions and a deeper understanding of the processes that drive solar flare activity, ultimately contributing to advancements in space weather forecasting.

6 Conclusion

In this work, we introduced a transformer-based model for predicting solar flares, employing the self-attention mechanism for the classification of Multivariate Time Series (MVTS) instances. Our study presents an innovative approach that harnesses the capabilities of the transformer model and the self-attention mechanism for MVTS classification. Through an end-to-end learning process, our proposed model effectively captures the temporal relationships inherent within MVTS instances. This includes the recognition of higher-order inter-variable relationships as well as local and global temporal changes. By incorporating attention-based techniques, our experiments conducted on a solar flare prediction dataset showcase the remarkable performance of our model in binary class MVTS classification, achieving an impressive TSS score of 70%. These outcomes underscore the potential of our approach to offer more comprehensive and precise predictions in the realm of solar physics and space weather forecasting. For future research, we intend to apply the Graph Attention Network [30] on the functional network constructed from the time series correlation so that the model can capture both spatial (inter-variable) and temporal dependencies for learning robust representations of the MVTS instances.

Acknowledgements This project has been supported in part by funding from the Division of Atmospheric and Geospace Sciences within the Directorate for Geosciences, under NSF awards #2301397, #2204363, and #2240022, and by funding from the Office of Advanced Cyberinfrastructure within the Directorate

for Computer and Information Science and Engineering, under NSF award #2305781. The authors acknowledge the use of ChatGPT (GPT-3.5) to rephrase sentences and improve the writing style of the manuscript.

References

1. Ahmadzadeh, A., Aydin, B., Georgoulis, M.K., Kempton, D.J., Mahajan, S.S., Angryk, R.A.: How to train your flare prediction model: Revisiting robust sampling of rare events. The Astrophysical Journal Supplement Series **254**(2), 23 (may 2021). https://doi.org/10.3847/1538-4365/abec88, https://dx.doi.org/10.3847/1538-4365/abec88
2. Ahmadzadeh, A., Aydin, B., Georgoulis, M.K., Kempton, D.J., Mahajan, S.S., Angryk, R.A.: How to train your flare prediction model: Revisiting robust sampling of rare events. The Astrophysical Journal Supplement Series **254**(2), 23 (May 2021). https://doi.org/10.3847/1538-4365/abec88, http://dx.doi.org/10.3847/1538-4365/abec88
3. Allouche, O., Tsoar, A., Kadmon, R.: Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (tss). Journal of Applied Ecology **43**(6), 1223–1232 (2006). https://doi.org/10.1111/j.1365-2664.2006.01214.x
4. Alshammari, K., Hamdi, S.M., Boubrahimi, S.F.: Feature selection from multivariate time series data: A case study of solar flare prediction. In: IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022. pp. 4796–4801. IEEE (2022). https://doi.org/10.1109/BIGDATA55660.2022.10020669, https://doi.org/10.1109/BigData55660.2022.10020669
5. Alshammari, K., Hamdi, S.M., Boubrahimi, S.F.: Identifying flare-indicative photospheric magnetic field parameters from multivariate time-series data of solar active regions. The Astrophysical Journal Supplement Series **271**(2), 39 (2024)
6. Alshammari, K., Hamdi, S.M., Muzaheed, A.A.M., Boubrahimi, S.F.: Forecasting multivariate time series of the magnetic field parameters of the solar events. CIKM workshop for Applied Machine Learning Methods for Time Series Forecasting (AMLTS) (2022)
7. Angryk, R.A., Martens, P.C., Aydin, B., Kempton, D., Mahajan, S.S., Basodi, S., Ahmadzadeh, A., Cai, X., Filali Boubrahimi, S., Hamdi, S.M., et al.: Multivariate time series dataset for space weather data analytics. Scientific data **7**(1), 1–13 (2020)
8. Bloomfield, D.S., Higgins, P.A., McAteer, R.T.J., Gallagher, P.T.: Toward reliable benchmarking of solar flare forecasting methods. The Astrophysical Journal Letters **747**(2), L41 (feb 2012). https://doi.org/10.1088/2041-8205/747/2/L41, https://dx.doi.org/10.1088/2041-8205/747/2/L41
9. Bobra, M.G., Couvidat, S.: Solar flare prediction using sdo/hmi vector magnetic field data with a machine-learning algorithm. The Astrophysical Journal **798**(2), 135 (2015)
10. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
11. Cui, Y., Li, R., Zhang, L., He, Y., Wang, H.: Correlation between solar flare productivity and photospheric magnetic field properties. Solar Physics **237**(1), 45–59 (2006)
12. Dempster, A., Schmidt, D.F., Webb, G.I.: MiniRocket: A very fast (almost) deterministic transform for time series classification. In: Proceedings of the 27th

ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 248–257. ACM, New York (2021)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
14. Eastwood, J., Biffis, E., Hapgood, M., Green, L., Bisi, M., Bentley, R., Wicks, R., McKinnell, L.A., Gibbs, M., Burnett, C.: The economic impact of space weather: Where do we stand?: The economic impact of space weather. Risk Analysis **37** (02 2017). https://doi.org/10.1111/risa.12765
15. Gao, J., Han, Y., Mao, Y.: A novel evaluation metric for imbalanced classification based on gini coefficient and tss. IEEE Access **8**, 80268–80280 (2020). https://doi.org/10.1109/ACCESS.2020.2996775
16. Hamdi, S.M., Aydin, B., Boubrahimi, S.F., Angryk, R., Krishnamurthy, L.C., Morris, R.: Biomarker detection from fmri-based complete functional connectivity networks. In: 2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). pp. 17–24. IEEE (2018)
17. Hamdi, S.M., Kempton, D., Ma, R., Boubrahimi, S.F., Angryk, R.A.: A time series classification-based approach for solar flare prediction. In: 2017 IEEE International Conference on Big Data (Big Data). pp. 2543–2551 (2017). https://doi.org/10.1109/BigData.2017.8258213
18. Hosseinzadeh, P., Boubrahimi, S.F., Hamdi, S.M.: Improving solar energetic particle event prediction through multivariate time series data augmentation. The Astrophysical Journal Supplement Series **270**(2), 31 (2024)
19. Hosseinzadeh, P., Filali Boubrahimi, S., Hamdi, S.M.: Toward enhanced prediction of high-impact solar energetic particle events using multimodal time series data fusion models. Space Weather **22**(6), e2024SW003982 (2024). https://doi.org/https://doi.org/10.1029/2024SW003982, `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024SW003982`, e2024SW003982 2024SW003982
20. Leka, K., Barnes, G.: Photospheric magnetic field properties of flaring versus flare-quiet active regions. ii. discriminant analysis. The Astrophysical Journal **595**(2), 1296 (2003)
21. Li, X., Zheng, Y., Wang, X., Wang, L.: Predicting solar flares using a novel deep convolutional neural network. The Astrophysical Journal **891**(1), 10 (2020)
22. McIntosh, P.S.: The classification of sunspot groups. Solar Physics **125**(2), 251–267 (1990)
23. Middlehurst, M., Large, J., Bagnall, A.J.: The canonical interval forest (CIF) classifier for time series classification. CoRR **abs/2008.09172** (2020), `https://arxiv.org/abs/2008.09172`
24. Muzaheed, A.A.M., Hamdi, S.M., Boubrahimi, S.F.: Sequence model-based end-to-end solar flare classification from multivariate time series data. In: 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021, Pasadena, CA, USA, December 13-16, 2021. pp. 435–440. IEEE (2021). https://doi.org/10.1109/ICMLA52953.2021.00074, `https://doi.org/10.1109/ICMLA52953.2021.00074`
25. Nguyen, T.L., Gsponer, S., Ilie, I., O'Reilly, M., Ifrim, G.: Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations. CoRR **abs/2006.01667** (2020), `https://arxiv.org/abs/2006.01667`
26. Nishizuka, N., Sugiura, K., Kubo, Y., Den, M., Watari, S., Ishii, M.: Solar flare prediction model with three machine-learning algorithms using ultraviolet brightening and vector magnetograms. Astrophysical Journal **835**(2), 156 (2017)

27. Saini, K., Alshammari, K., Hamdi, S.M., Filali Boubrahimi, S.: Classification of major solar flares from extremely imbalanced multivariate time series data using minimally random convolutional kernel transform. Universe **10**(6) (2024). https://doi.org/10.3390/universe10060234, https://www.mdpi.com/2218-1997/10/6/234
28. Song, H., Tan, C., Jing, J., Wang, H., Yurchyshyn, V., Abramenko, V.: Statistical assessment of photospheric magnetic features in imminent solar flare predictions. Solar Physics **254**(1), 101–125 (2009)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
30. Velivckovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. In: Proceedings of the 5th International Conference on Learning Representations (ICLR) (2018)
31. Yu, D., Huang, X., Wang, H., Cui, Y.: Short-term solar flare prediction using a sequential supervised learning method. Solar Physics **255**(1), 91–105 (2009)
32. Zheng, X., Zhang, C., Woodland, P.C.: Adapting gpt, gpt-2 and bert language models for speech recognition (2021)
33. Zheng, Y., Li, X., Wang, X.: Solar flare prediction with the hybrid deep convolutional neural network. The Astrophysical Journal **885**(1), 73 (2019)