Using a Transformer Architecture to Detect Stellar Flares in TESS Lightcurves

by

Isabella Longo

A thesis submitted to the

Faculty at the University of Colorado in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Department of Computer Science

2025

Committee Members:

Liz Bradley, Chair

Rachel Cox

Majid Zamani

David Wilson

Longo, Isabella (Computer Science)

Thesis directed by Professor Liz Bradley

This thesis introduces a novel method for detecting stellar flares in light curve data obtained from the Transiting Exoplanet Survey Satellite (TESS) by employing a transformer-based deep learning architecture. Stellar flares—sudden and explosive releases of energy resulting from magnetic field reconnection on stellar surfaces—are significant phenomena that offer valuable insights into stellar activity and may influence the habitability of orbiting exoplanets. Traditional detection techniques, such as threshold-based methods, template fitting, and convolutional neural networks, often struggle to identify flares amid the complex variability of background stellar activity.

The transformer model utilizes self-attention mechanisms to effectively capture temporal relationships within light curve data, allowing for the identification of flare signatures regardless of their position in the observation window. Trained on a dataset of approximately $53,000$ labeled flare events from the Pietras catalog, the model operates on fixed-length windows of 100 time steps, with specific architectural design choices guided by thorough hyperparameter optimization studies. Experimental results indicate that the transformer surpasses existing methods, achieving an F1 score of 0.83, compared to 0.82 for CNN-based approaches and 0.78 for threshold methods. The model demonstrates notable advantages in challenging scenarios, such as enhancing the detection of low-energy flares (with a recall of 0.77 for flares below 10^{32} erg versus 0.70 for the STELLA-like CNN approach), maintaining consistent performance across various stellar variability patterns, and accurately identifying multiple flares within single observation windows (achieving an F1 score of 0.85 compared to 0.77 for CNNs). Attention visualizations provide interpretable insights into the model's decision-making process, revealing focused attention on flare rise and early decay phases while comparing these features with the surrounding baseline flux. Error analysis identifies instrumental artifacts and cosmic ray hits as primary sources of false positives, while low-amplitude flares constitute the majority of false negatives.

This research demonstrates the potential of transformer architectures to advance the automated detection of astrophysical events in time series data. It contributes to more comprehensive catalogs of stellar activity, which has implications for understanding stellar magnetic evolution and exoplanet habitability.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Detecting stellar flares in light curve data is a complex task that requires distinguishing transient astrophysical events from background variability and noise. This challenge lies at the intersection of astrophysics and machine learning, where novel computational methods can enhance detection accuracy. This thesis proposes a transformer-based approach to identify these energetic stellar events with higher accuracy and precision than existing methods, reducing the high rate of false positives in existing models. The following sections provide context for this research, outlining the nature of stellar flares, their significance in astrophysics, current detection challenges, and how transformer architectures offer promising solutions.

1.1 Nature and Significance of Stellar Flares

Stellar flares are intense, sudden outbursts of energy on the surfaces of stars caused by the reconnection of magnetic fields [5]. These events occur when intense magnetic fields on the star become tangled and snap, releasing substantial energy across the electromagnetic spectrum. Flares manifest as rapid increases in brightness followed by a more gradual decay phase as the star returns to its quiescent state. These energetic phenomena provide crucial insights into stellar magnetic activity and evolutionary processes that would otherwise remain unobservable from Earth.

Space-based telescopes such as the Transiting Exoplanet Survey Satellite (TESS) have revolutionized the detection and analysis of stellar flares. TESS monitors the brightness of stars across the sky, capturing continuous light curves that record minute variations in stellar flux over time. A light curve represents the brightness of a star over time, where flares appear as distinct, sharp spikes of increased luminosity against the background of normal stellar variability.

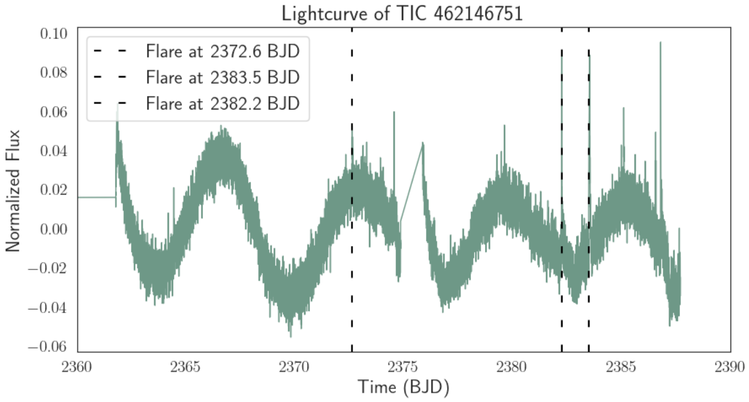

Figure 1.1: Example light curve from TESS data showing stellar flares. The sharp upward spikes indicate flare events, characterized by sudden increases in brightness followed by gradual decay. The underlying quasi-periodic variations represent normal stellar activity such as rotation, which complicate flare detection.

Studying stellar flares extends beyond pure stellar astrophysics, significantly impacting exoplanetary research. Flares can profoundly impact planets orbiting the host star in several critical ways:

(1) Large flares can strip away a planet's atmosphere through increased radiation and particle flux, potentially rendering otherwise habitable planets uninhabitable [33].

(2) Flares can alter the chemical composition of planetary atmospheres, creating permanent changes to environmental conditions [33].

(3) In binary star systems, flares from one star can affect the orbit and evolution of companion stars, influencing mass transfer rates and system stability [22].

Understanding flare frequency, intensity, and characteristics thus becomes essential when assessing potential habitability zones and the long-term evolution of planetary systems.

1.2 Current Detection Challenges

Despite their importance, detecting stellar flares in TESS light curves presents significant challenges. TESS observes millions of light curves across 26 sectors, generating an immense volume of data that makes manual identification impractical. Stellar light curves also exhibit complex patterns of variability due to factors such as stellar rotation, spots, and instrumental noise, making it difficult to distinguish genuine flare events from other brightness fluctuations. As shown in Figure 1.1, flares appear as sharp upward spikes against backgrounds that may include quasi-periodic variations and other stellar activity, complicating automated detection approaches.

There are three main methods currently used for automated stellar flare detection, each with distinct limitations:

- **Stella**: While this Convolutional Neural Network (CNN) approach automates flare detection, CNNs are fundamentally limited in their ability to capture time-series data patterns effectively. CNNs excel at spatial feature extraction but struggle with the temporal dependencies that characterize flare events [17].

- **PyVAN**: This approach assesses stellar flare candidate light curves through template optimization but faces difficulties when confronted with unclear signals. Its reliance on empirical templates limits precision when distinguishing flares from other transient events, particularly in noisy data [34].

- **Celerite QFD**: This method uses hidden Markov Models (HMMs) for flare detection and is particularly sensitive to noise in photometric data. The rigid state-based assumptions of HMMs constrain the model to a predefined set of discrete states (Quiet, Firing, and Decaying), which cannot adequately represent the continuous spectrum of flare morphologies observed in stellar light curves [16].

While useful, these methods struggle to generalize across different stellar types and noise conditions, necessitating an adaptable and efficient approach to learning temporal dependencies.

1.3 Transformer as a Solution

Transformer architectures represent a potentially novel approach to flare detection problems. Initially developed for natural language processing, transformers have demonstrated exceptional capabilities in modeling sequential data through their self-attention mechanisms [58]. Unlike recurrent neural networks or CNNs, transformers can dynamically attend to relevant parts of the input sequence regardless of their position. The self-attention mechanism allows transformers to weigh the importance of different time steps in the input sequence relative to each other. For flare detection, the model can learn to focus on the distinctive patterns of rapid brightness increase followed by exponential decay that characterize flare events while considering the broader context of the star's behavior. Additionally, the transformer's parallel processing capabilities enable efficient handling of large datasets, making them computationally tractable for the extensive TESS observations.

Several key advantages make transformers particularly well-suited for stellar flare detection:

- **Effective Temporal Pattern Recognition**: Transformers excel at capturing both short-term and long-term dependencies in sequential data, which is essential for distinguishing flare signatures from background noise and variability.
- **Parallel Processing**: Unlike recurrent models that process data sequentially, transformers analyze the entire sequence simultaneously, significantly increasing computational efficiency when dealing with extensive light curve datasets [58].
- **Contextual Understanding**: Through multi-head attention mechanisms, transformers can simultaneously focus on different aspects of the light curve. This enables them to consider both local intensity changes and broader patterns of stellar behavior [35]

1.4 Approach

This thesis proposes a transformer-based model for detecting stellar flares in TESS light curve data. The research utilizes a comprehensive dataset of approximately 53,000 flare events from the catalog compiled by Pietras et al. (2022) [46], providing a robust foundation for model training and evaluation.

The approach involves preprocessing light curve data to create fixed-length windows centered around potential flare events, each containing 100-time steps. These windows serve as input to the transformer model, which learns to distinguish genuine flares from false positives by attending to the temporal patterns that characterize flare morphology. The model's architecture includes multiple self-attention layers, positional encodings to preserve temporal information, and a classification head that outputs the probability of a flare at each time step.

By leveraging mechanisms, this research aims to reduce false positive rates and improve the recall of low-energy flares, leading to a more reliable catalog of stellar activity events. It aims to contribute to understanding stellar activity and its implications for planetary systems.

Chapter 2

Background

This chapter provides the foundation for understanding stellar flares, current detection methodologies, and the application of transformer architectures to time-series data analysis. Beginning with the physical processes that generate stellar flares, this section describes the TESS mission that supplies the light curve data central to this research. The discussion then examines current flare detection algorithms, highlighting their strengths and limitations before introducing transformer architectures, and their potential advantages for this particular application.

2.1 Stellar Flare Physics

Stellar flares represent one of the most energetic manifestations of magnetic activity on stars. These explosive events occur when magnetic energy stored in a star's atmosphere is suddenly released through magnetic reconnection [5]. The typical structure of a stellar magnetic field includes closed loops that extend from the stellar surface into the corona. Over time, these magnetic field lines can become twisted and stressed due to the differential rotation of the star and convective motions in its outer layers [53].

When the stress in these magnetic fields reaches a critical threshold, reconnection occurs—the topology of the magnetic field lines rapidly reconfigures, converting magnetic energy into thermal energy, kinetic energy, and electromagnetic radiation [18]. This process accelerates charged particles to relativistic speeds and heats plasma to temperatures exceeding 10 million Kelvin[20]. The resulting emission spans across the electromagnetic spectrum, from radio waves to X-rays and gamma rays, though ground-based observations are typically limited to optical wavelengths.

The temporal profile of a stellar flare typically follows a characteristic pattern: a rapid rise in brightness (the impulsive phase) followed by a more gradual decay (the gradual phase) [12]. The impulsive phase, lasting from seconds to minutes, represents the initial energy release and particle acceleration. The gradual phase, which can extend from minutes to hours, reflects the cooling of heated plasma and the relaxation of the magnetic field structure. This distinctive temporal signature—sharp rise and exponential decay as shown in Figure 2.1—provides the primary basis for scientists detecting flares in light curve data by hand.

The energy released during stellar flares varies enormously, spanning several orders of magnitude. On our Sun, typical flares release energy equivalent to 10^{20} to 10^{25} joules [51]. However, more active stars, particularly M-dwarfs, can produce "superflares" with energies up to 10^{33} joules—thousands of times more energetic than the strongest solar flares recorded. The frequency of flares typically follows a power-law distribution, with more minor flares occurring much more frequently than larger ones [3].

Understanding the physical mechanisms behind stellar flares provides a crucial context for developing detection algorithms. Flares' characteristic temporal evolution and spectral properties inform the design of features and architectures that can effectively distinguish these events from other forms of stellar variability and instrumental noise.

2.2 TESS Mission and Light Curve Data

TESS, launched by NASA in April 2018, represents a significant advancement in our capability to monitor stellar activity and detect exoplanets. While primarily designed to identify exoplanets through the

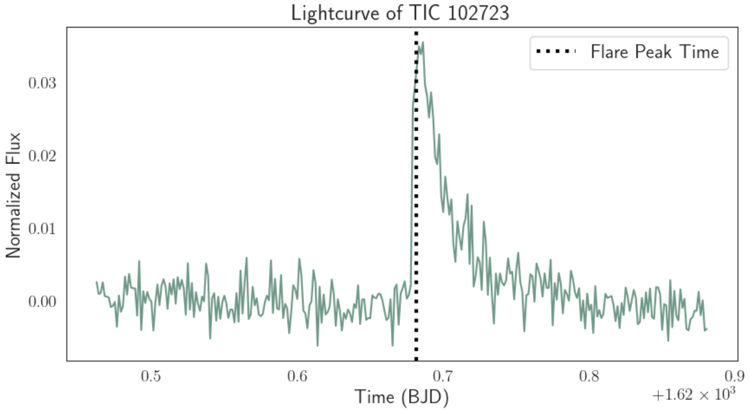

Figure 2.1: Example of a stellar flare from TESS light curve data for TIC 102723. The light curve shows the characteristic flare profile with a rapid rise followed by a more gradual decay. The dotted line indicates the flare peak time. This pattern—sharp rise and exponential decay—provides the primary basis for detecting flares in light curve data. The x-axis represents time in Barycentric Julian Date (BJD), and the y-axis shows the normalized flux, where the star's quiescent state is normalized to zero.

transit method, TESS provides unprecedented data for studying stellar variability, including flares, across a large sample of stars.

TESS, orbiting in a elliptical high-Earth orbit, observes the sky in sectors, each covering a 24° × 96° region and being monitored continuously for approximately 27 days [56]. The satellite's four cameras capture images every two seconds, which are then combined into two-minute cadence data for selected targets and 30-minute cadence data for full-frame images. This high-cadence, long-duration monitoring makes TESS particularly well-suited for detecting and characterizing stellar flares, which can occur on timescales of minutes to hours.

A light curve derived from TESS data represents the brightness of a star measured over time, typically normalized to the star's median flux. These measurements are precise enough to detect brightness variations as small as 0.01%, allowing for even relatively modest flare events. However, TESS light curves also capture other forms of stellar variability, including:

- Rotational modulation due to starspots
- Pulsations in variable stars
- Binary eclipses

Distinguishing flares from these other phenomena poses a significant challenge for several reasons. While flares are characterized by their rapid rise and exponential decay profile, other types of stellar variability can create confusion. Rotational modulation occurs as starspots rotate in and out of view, creating quasiperiodic brightness variations that establish an irregular baseline, making it difficult to determine the pre-flare quiescent level necessary for accurate flare detection [21]. Stellar pulsations present another challenge, as some variable stars exhibit brightness fluctuations on timescales similar to the duration of flare events (minutes to hours), and these pulsations can sometimes produce sharp brightness increases that mimic the impulsive phase of flares [64]. Binary eclipses contribute to detection difficulties when the ingress or egress phases of eclipsing binary systems produce relatively rapid changes in brightness that, when combined with noise, might be misidentified as flare events. Additionally, instrumental effects such as momentum dumps, cosmic ray hits, and other systematic effects in the TESS data can produce transient brightness changes that superficially resemble flare signatures [32]. The baseline flux level may be variable due to multiple overlapping processes co-occurring, further complicating the identification of anomalous brightening events against this complex background.

The sheer volume of TESS data further compounds these challenges. With observations of approximately 200,000 selected targets at two-minute cadence across its primary mission [55], plus millions more stars in full-frame images, manual inspection becomes impractical. This necessitates automated methods for detecting and characterizing flares, driving the development of algorithms that can efficiently process large datasets while maintaining high accuracy and low false-positive rates.

2.3 Current Flare Detection Methods

Several approaches have been developed to automate the detection of stellar flares in light curve data. These methods vary in their underlying principles, computational requirements, and effectiveness across different types of stellar variability.

2.3.1 Threshold-Based Methods

The most straightforward and traditional approach involves setting amplitude thresholds above the local mean or median flux level. Events exceeding these thresholds by a specified number of standard deviations are flagged as potential flares [44]. While straightforward, these methods often struggle with variable backgrounds and complex noise profiles. Additionally, they typically require careful parameter tuning for each light curve, limiting their scalability to large datasets like those provided by TESS.

2.3.2 Model Fitting Approaches

More sophisticated methods employ model fitting to characterize flares based on their expected temporal evolution. For example, the widely used FBEYE algorithm fits a flare template to potential events, allowing for the extraction of parameters such as amplitude, duration, and energy [13]. PyVAN extends this approach by employing the Differential Evolution algorithm to optimize empirically derived lightcurve templates for various types of stars. This evolutionary optimization approach explores parameter space to identify the template configuration that best explains observed brightness variations. These model-based approaches can provide detailed characterization of flare events. However, they may fail to generalize across diverse stellar types if the template models do not adequately capture the full range of flare morphologies.

2.3.3 Machine Learning Models

In recent years, the application of machine learning techniques to flare detection has increased. The STELLA model developed by Feinstein et al. (2020) employs a convolutional neural network (CNN) to identify flares in TESS data [17]. This approach automates feature extraction and can learn complex patterns that might be missed by threshold or template-based methods. However, CNNs are primarily designed for spatial pattern recognition and may not fully capture the temporal dependencies that characterize flare events in time series data [24].

Another machine learning approach, celeriteQFD, combines hidden Markov models (HMMs) with a celerite process, a computational technique for Gaussian Process regression specifically optimized for one-dimensional datasets with irregular sampling intervals, to model stellar flares as discrete states: Quiet, Firing, and Decaying [16]. While effective at detecting flares in variable and noisy light curves, this method's performance may suffer from rigid state-based assumptions and sensitivity to noise, limiting its ability to generalize across different stellar types and flare shapes.

2.3.4 Limitations of Current Approaches

Despite advances in flare detection methodology, several limitations persist across current approaches:

- **Background Variability**: Many methods struggle to accurately establish the baseline flux level against which flares are detected, particularly for stars with significant rotational modulation or other forms of variability [64].

- **Noise Sensitivity**: Both threshold-based and model-fitting approaches are susceptible to false positives from instrumental noise, cosmic rays, or other transient events that may mimic flare signatures [32].
- **Scalability**: Methods requiring extensive parameter tuning or significant computational resources may not scale effectively to the large datasets produced by missions like TESS [59].
- **Generalization**: Techniques optimized for specific types of stars or flare morphologies may not perform well across the diverse stellar population observed by TESS [22].
- **Temporal Dynamics**: Many approaches, particularly CNN-based methods, do not fully exploit the temporal structure of light curves, treating each time point somewhat independently rather than as part of a coherent sequence [15].

These limitations motivate the exploration of alternative architectures that can better address the temporal nature of flare detection while remaining robust to variability and noise. The transformer architecture, with its ability to model long-range dependencies in sequential data, offers a promising solution to these challenges.

2.4 Introduction to the Transformer

Transformer architectures represent a paradigm shift in sequence modeling, introduced by Vaswani et al. in 2017 in their seminal paper "Attention Is All You Need" [58]. Initially developed for natural language processing tasks, transformers have demonstrated capabilities across numerous domains involving sequential data, including time series analysis [66].

2.4.1 Core Components

The transformer's architecture consists of several key components that distinguish it from previous sequence modeling approaches:

- **Self-Attention Mechanism**: At the heart of the transformer is the self-attention mechanism, which allows the model to weigh the importance of different elements in a sequence relative to each other. This contrasts with recurrent neural networks (RNNs), which process sequences step by step, and CNNs, which operate on fixed local regions through convolutional filters.
- **Multi-Head Attention**: Transformers typically employ multiple attention "heads" in parallel, each learning to focus on different aspects of the relationships between sequence elements. This enables the model to simultaneously capture various types of dependencies and patterns within the data
- **Positional Encoding**: Since the self-attention mechanism itself is permutation-invariant (it does not inherently consider the order of elements), transformers incorporate positional encodings to inject information about the position of each element in the sequence. These encodings are added to the input embeddings, allowing the model to leverage content and position information.
- **Feed-Forward Networks**: Each transformer layer includes point-wise feed-forward networks that process the output of the attention mechanism, adding non-linearity and representational capacity to the model.
- **Layer Normalization and Residual Connections**: These components help stabilize training and allow for the construction of deep transformer networks by mitigating issues such as vanishing gradients.

2.4.2 Advantages of Sequential Data Processing

Several characteristics make transformers particularly well-suited for processing sequential data like stellar light curves:

- **Parallel Processing**: Unlike RNNs, which process sequences step by step, transformers simultaneously compute attention for all sequence elements. This parallelization enables more efficient training and inference, especially for long sequences.
- **Long-Range Dependencies**: The self-attention mechanism allows transformers to capture dependencies between any pair of sequence elements, regardless of their positional distance within the time series. This contrasts with CNNs, which are limited by their receptive field size, and RNNs, which often struggle with long-range dependencies due to vanishing gradients.
- **Flexible Context Integration**: Transformers can dynamically adjust how much context from different parts of the sequence influences the representation of each element. This adaptability is valuable for time series data where the relevant context may span varying temporal scales.

2.5 Applications of Transformers in Time-Series Analysis

While initially developed for natural language processing, transformers have been successfully adapted to various time series analysis tasks, demonstrating their versatility and effectiveness in this domain.

2.5.1 Anomaly Detection

Transformers' ability to model standard patterns in sequential data makes them practical tools for anomaly detection. Approaches such as the Time Series Transformer (TST) [63] learn representations of normal behavior and identify deviations that could indicate anomalies. This capability aligns well with flare detection, which can be framed as identifying anomalous brightness increases in stellar light curves.

2.5.2 Classification Tasks

Transformer-based architectures have demonstrated strong performance in time series classification tasks across domains such as healthcare, finance, and industrial monitoring. Their ability to attend to relevant portions of the input sequence while ignoring irrelevant or noisy segments makes them robust to the variability and noise often present in real-world time series data [66].

2.5.3 Astronomical Applications

In astronomy, transformers have begun to find applications in various time-domain problems. Salinas et al. (2023) utilized a transformer-based approach for distinguishing planetary transits from false positives in light curve data [50]. Similarly, Jadhav et al. (2021) applied transformer architectures to classify variable stars based on their light curves. These successful applications suggest that transformers can effectively capture the complex patterns in astronomical time series data.

The demonstrated success of transformers across diverse time series applications, combined with their inherent advantages for sequential data processing, positions them as promising candidates for stellar flare detection. By leveraging their ability to capture local and global patterns in light curves, transformer-based models may overcome many limitations associated with current detection methods.

Chapter 3

Methodology

This chapter presents a comprehensive methodology for detecting stellar flares using a transformer-based architecture. The chapter encompasses data acquisition and preprocessing, model architecture design, and implementation details. By leveraging the transformer's self-attention mechanism, this research aims to classify the temporal dynamics of flare events more effectively than existing methods, reducing false positive rates while improving the detection of low-energy flares.

3.1 Dataset Preparation and Preprocessing

3.1.1 Data Acquisition

The primary dataset for this research comes from the catalog compiled by Pietras et al. (2022) [46], which contains over 147,000 stellar flares identified on more than 25,000 stars across TESS sectors 1-39. This catalog represents one of the most extensive collections of TESS flare events, providing a robust foundation for model training and evaluation. The catalog includes crucial parameters for each flare, including TIC identifier, TESS sector, flare peak time in Barycentric Julian Date (BJD), flare amplitude, duration, number of fitted flare profiles, equivalent duration, and flare energy estimates.

The Pietras catalog was developed using their custom WARPFINDER (Wroclaw AlgoRithm Prepared For detectINg anD analyzing stEllar flaRes) software, which employs a three-step methodology for flare detection. First, a trend method identifies potential flares by detrending light curves with multiple smoothing windows and looking for points that protrude above the average trend. Second, a difference method examines the differences between consecutive points, marking points that exceed 3σ as potential flare detections. Finally, a flare profile method fits mathematical models to the potential flares, using single-profile and double-profile fits to capture different flare morphologies. The Pietras team found that approximately 7.7% of all stars in their sample exhibited flares, over 50% for M-type stars.

What makes this catalog particularly valuable is its thorough approach to validation. The automated detection was supplemented by visual inspection to eliminate false positives, especially prevalent in stars of spectral types earlier than F0. The catalog spans a diverse range of stellar types, with flare energies ranging from 10^{31} to 10^{36} erg, ensuring that the model is exposed to flares with varying morphologies, durations, and intensities.

For non-flare examples, this research utilized the stars from the Pietras catalog, classified as having zero flares. The Pietras team analyzed approximately $330,000$ stars across 39 TESS sectors but only found flaring activity in about $25,000$ stars (7.7% of the sample). This provided a substantial pool of approximately $305,000$ non-flaring stars to draw negative examples. Using non-flaring stars from the same catalog ensured consistency in the data processing pipeline and observation characteristics, which is critical for developing unbiased training dataset.

3.1.2 LightCurve Extraction and Processing

The light curves from TESS data were extracted using the `lightkurve` Python package [36], which provides tools for accessing and manipulating time series data from space-based photometric missions. For each cataloged flare, a query was made to the MAST (Mikulski Archive for Space Telescopes) to retrieve the corresponding TESS light curve.

The light curve extraction process involved several sequential operations. First, the TESS database was queried for the target star using its TESS Input Catalog (TIC) identifier and sector information. Next, the two-minute cadence light curve data was downloaded from the archive. The data was then processed by removing NaN values and 5σ outliers to ensure data quality. Then, flux values were normalized by dividing by the median flux and subtracting 1, resulting in a relative flux where 0 represents the quiescent state. Finally, the light curves were appended for stars with observations across multiple sectors to create a longer record where possible, providing more extended baseline observations.

A critical decision point in the methodology concerned whether to remove the underlying quasiperiodic variability from the light curves before flare detection. Unlike previous approaches that detrend the light curves to produce a flat baseline, this research maintains the natural variability of the stellar flux. This decision was made based on preliminary experiments showing that the transformer architecture could effectively learn to distinguish flares from other types of variability without explicit detrending. Maintaining the original variability also preserves important contextual information about the stellar environment in which flares occur. This allows the model to identify patterns associated with flare probability that might be lost in detrended data.

This approach differs from traditional methods that often require detrending to establish a clear baseline for threshold-based detection. It highlights one of the transformer's advantages: learning complex patterns directly from raw data. However, to ensure the model focuses on the relevant features, the flux values were standardized using min-max normalization within each window to constrain the input range.

3.1.3 Window Extraction and Data Augmentation

A fixed-length window of 100-time steps was extracted for each identified flare in the catalog, centered on the flare peak time with some random offset to prevent position bias. The window size of 100-time steps (approximately 200 minutes at the TESS two-minute cadence) was chosen to include that data before and after the flare event while keeping the sequence length manageable for the transformer model.

To address the challenge of multiple flares occurring within a single observation window, an additional processing step was implemented to create a mask indicating the positions of all flares within each window. This approach allows the model to learn from examples with multiple flares without confusing them as false positives or negatives. The methodology for identifying multiple flares involved a systematic approach to capture all flare events within a single observation window. Other cataloged flares from the same star were checked for each primary flare to determine if they occurred within the extracted time window. A binary mask array of length 100 was then created, where a value of 1 indicated the presence of a flare at that specific time step. This mask was included as additional information during the training process to help the model distinguish between single and multiple flare scenarios, allowing it to learn the complex patterns that might emerge when multiple energetic events occur in close temporal proximity.

Data augmentation techniques enhanced the model's robustness and generalization capability. The augmentation strategy incorporated several complementary approaches. Random temporal shifting was applied so that the position of the flare peak within the window was randomly varied rather than fixed at the center position, forcing the model to identify flares regardless of their position in the sequence. Amplitude scaling was utilized where flare amplitudes were randomly scaled within a small range (±10%) to help the model generalize across different flare intensities. Additionally, small amounts of additive Gaussian noise were introduced to simulate instrumental variations and improve robustness. Some examples were also time-reversed to prevent the model from learning directional biases that might not reflect the physical reality of flare detection. These augmentation techniques are standard in time-series deep learning pipelines and have been widely adopted to improve model generalization and reduce overfitting [66]].

3.1.4 Computational Infrastructure for Dataset Preparation

The dataset preparation process was computationally intensive, mainly due to the need to query and download light curves for thousands of stars. This processing was performed on CU Boulder's Alpine supercomputing cluster. I implemented a parallelized querying system using Python's `ProcessPoolExecutor`

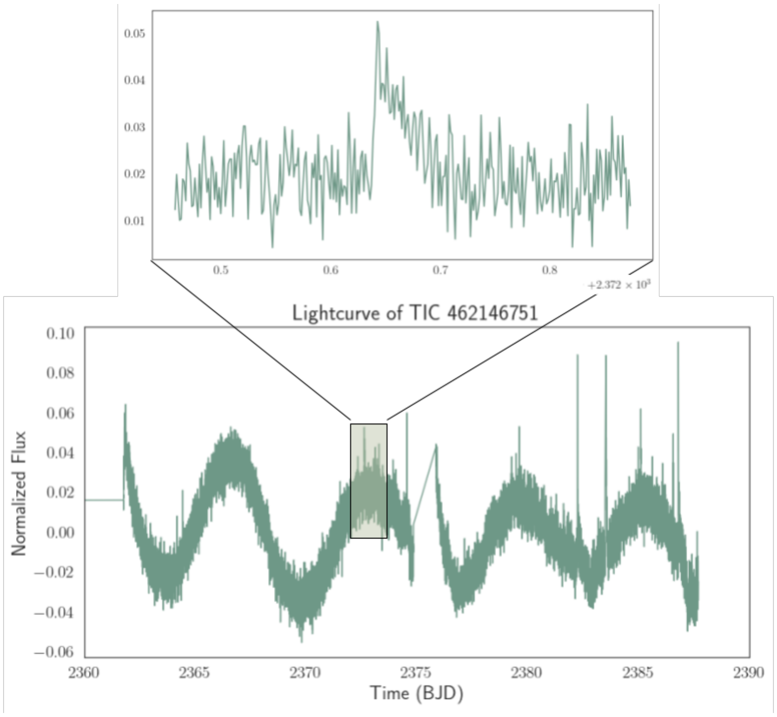

Figure 3.1: Example of a stellar flare from TESS light curve data for TIC 462146751. The light curve shows the characteristic flare profile with a rapid rise followed by a more gradual decay. This pattern provides the primary basis for detecting flares in light curve data by hand. The x-axis represents time in Barycentric Julian Date (BJD), and the y-axis shows the normalized flux.

from the concurrent to optimize the data acquisition `process.futures` module, efficiently distributing query tasks across available CPU cores. The pipeline was designed to run on 2 CPU cores with workloads divided into 12-hour processing sets. This parallelization strategy significantly reduced the otherwise sequential download time. Still, it required approximately 5 days total to complete due to rate limits on the MAST archive and the sheer volume of data. The final processing and training steps leveraged an NVIDIA RTX 8000 GPU with 48GB of VRAM, essential for handling the memory-intensive transformer architecture. The entire data preparation pipeline, including light curve downloading, window extraction, and augmentation, was executed in multiple stages with checkpointing to ensure resilience against potential job interruptions on the shared computing cluster.

The final dataset was organized into training (80%), validation (10%), and test (10%) splits, ensuring that all examples from the same star were kept in the same split to prevent data leakage. The dataset was stored in HDF5 format, which provides efficient storage and access for large numerical datasets. The training set contained approximately $42,400$ flare windows and $80,000$ non-flare windows, while the validation and test set each contained about $5,300$ flare windows and $10,000$ non-flare windows.

3.2 Transformer Architecture

3.2.1 Motivation and Mathematical Foundation

The transformer architecture was selected for this task due to its proven capabilities in modeling long-range dependencies in sequential data. While recurrent architectures like Long Short-Term Memory (LSTM) networks have been widely used for time series tasks, they face limitations in capturing long-range dependencies due to the vanishing gradient problem. Additionally, their sequential processing nature limits parallelization during training. While parallelizable, convolutional Neural Networks (CNNs) are limited by their fixed receptive field and struggle to capture global patterns across the entire sequence.

Transformers address these limitations through their self-attention mechanism, which computes weighted relationships between all pairs of positions in the input sequence. This allows the model to directly capture dependencies between any two-time steps, regardless of their distance in the sequence. The self-attention mechanism is defined mathematically as:

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

Where:

- Q (query), K (key), and V (value) are different linear projections of the input sequence
- d_k is the dimensionality of the key vectors
- The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the softmax function from entering regions with minimal gradients

In flare detection, the model can compare a given time step with all other time steps in the window, identifying patterns such as flare events' characteristic sharp rise and exponential decay against the background of stellar variability.

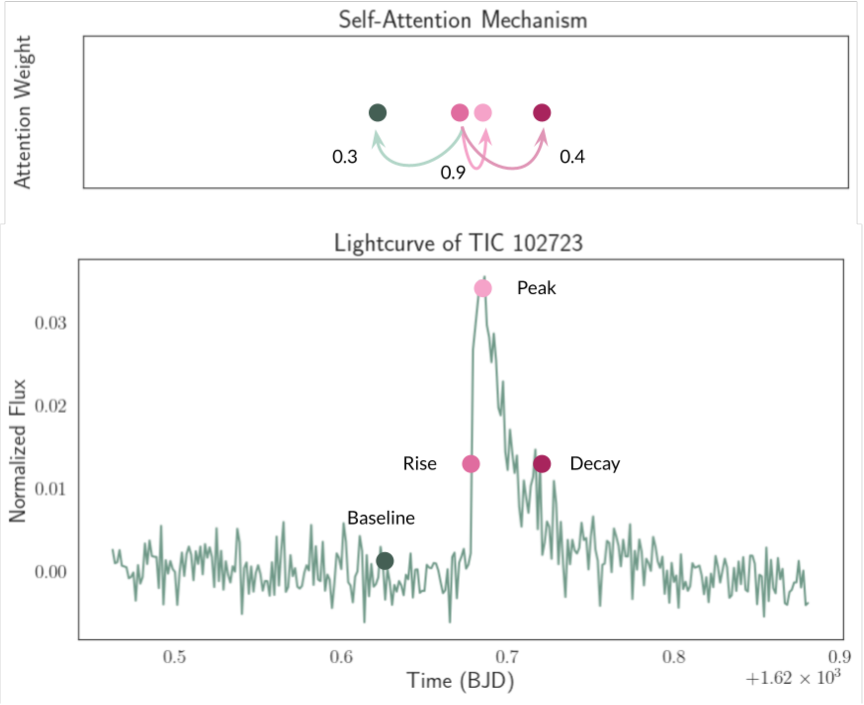

Figure 3.2: Visualization of the self-attention mechanism applied to a stellar flare light curve from TIC 102723. The top panel demonstrates how self-attention assigns weights to different parts of the sequence, with the rise phase strongly attending to the peak (0.9 weight) and decay phase (0.4 weight), while giving less attention to the baseline (0.3 weight). The bottom panel shows the corresponding light curve with key phases labeled: baseline (green), rise (pink), peak (light pink), and decay (burgundy). This illustrates how the transformer model can dynamically focus on the most relevant portions of the light curve for flare detection, capturing the temporal relationships between different phases of the flare event.

The transformer employs multi-head attention, which performs the attention operation in parallel across multiple representation subspaces:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, ..., \text{head}_h)W^O$$

Where:

- $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$
- The W matrices are learnable parameter matrices
- h is the number of attention heads

This multi-head approach allows the model to jointly attend to information from different representation subspaces and positions, providing a more comprehensive view of the temporal dynamics in the light curve.

3.2.2 Model Architecture Details

The transformer model for flare detection consists of an encoder-only architecture similar to BERT [14], as the task requires understanding the input sequence but not generating a new sequence. The model processes fixed-length windows of light curve data and outputs a binary classification indicating the presence or absence of a flare. The architecture includes the following components:

(1) **Input Embedding Layer**: Transforms the input features (primarily flux values) into a higher-dimensional representation space.

(2) **Positional Encoding**: Since the transformer does not understand sequence order, positional encodings are added to the input embeddings to provide information about each time step's relative or absolute position in the sequence. The positional encodings use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

Where pos is the position in the sequence, i is the dimension, and d is the embedding dimension.

(3) **Encoder Blocks**: The core of the transformer architecture, consisting of multi-head self-attention layers followed by position-wise feed-forward networks. Each encoder block also includes residual connections to preserve gradient flow, and layer normalization to standardize activations across features, enhancing training stability [58]:

(a) Multi-Head Self-Attention Block:

$$\text{LayerNorm}(x + \text{MultiHeadAttention}(x))$$

(b) Feed-Forward Network Block:

$$\text{LayerNorm}(x + \text{FeedForward}(x))$$

The implementation uses two encoder blocks with eight attention heads each. The dimensionality of the model (d) was set to 256, and the feed-forward network dimensionality to 512. These values were selected based on experiments on the commonly used configurations in transformer models for time series data, balancing performance and computational cost [66], [58].

(4) **Classification Head**: Processes the encoded sequence to produce a final prediction. The implementation includes:

- A flattening operation to convert the sequence representation to a vector
- A dense layer with ReLU activation (512 units)
- Another dense layer with ReLU activation (128 units)
- A final output layer with sigmoid activation for binary classification

The model's hyperparameters were determined through experiments on the validation set, as well as utilizing standard hyperparameters for a small transformer model, detailed in Section 3.4. Key hyperparameters included an embedding dimension of 256, two encoder blocks, eight attention heads, a feed-forward network dimension of 512, a dropout rate of 0.1 (applied after each sub-layer to prevent overfitting), and a maximum sequence length of 100-time steps.

The relatively moderate complexity of the model (in comparison to large language models that use transformers) was chosen to balance performance against computational efficiency, recognizing that the light curve data, while complex, has fewer dimensions and simpler structures than natural language processing tasks.

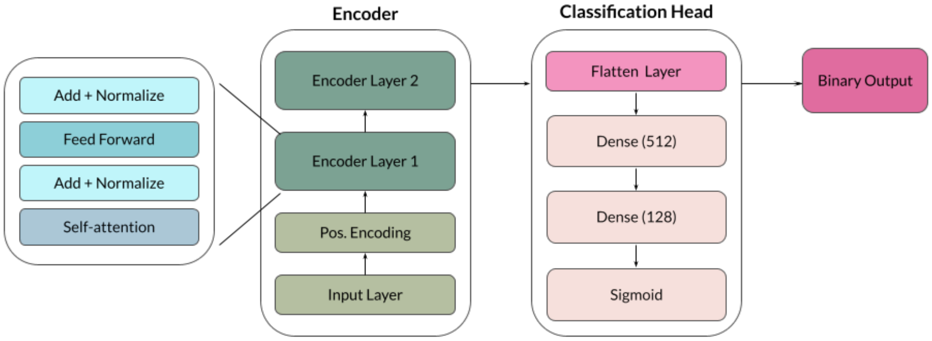

Figure 3.3: Architecture of the transformer model for stellar flare detection. The model consists of three main components: (1) Encoder blocks with positional encoding, input layer, and stacked encoder layers, where each encoder layer contains self-attention mechanisms, add and normalize operations, and feed-forward networks; (2) Classification head with a flatten layer followed by dense layers (512 and 128 units) and a sigmoid activation; and (3) Binary output for flare classification. This architecture enables the model to capture complex temporal patterns in light curve data while maintaining sequence information through positional encoding.

3.2.3 Input and Output Representation

The input to the transformer model consists of windows of light curve data, each containing 100 time steps. For each time step, several features were provided: the normalized flux value, local flux gradient (calculated as the difference between consecutive flux values), and a binary indicator for potential data quality issues (e.g., momentum dumps, spacecraft maneuvers).

These features were combined to form a 3-dimensional input vector for each time step. The sequence of 100 vectors was then processed by the transformer to produce a single binary output indicating whether the window contains a flare. In windows with multiple flares, the model was trained to output a positive result (1) if any flare was present.

The additional flare position mask described in Section 3.1.3 was used as an auxiliary input for training examples where multiple flares occurred within the same window. This allowed the model to learn that multiple brightness spikes within a single window could all represent genuine flares rather than false positives.

3.3 Implementation

The transformer model was implemented using `PyTorch` [45], a flexible deep-learning framework that provides efficient tensor computations and automatic differentiation capabilities. `PyTorch` was selected over alternatives like `TensorFlow` due to its dynamic computational graph, which facilitates rapid experimentation and debugging during model development. The implementation leveraged PyTorch's native transformer modules (`nn.TransformerEncoderLayer` and `nn.TransformerEncoder`) as foundation components, with custom modifications to adapt the architecture to the flare detection task.

The codebase was structured following a modular, object-oriented design pattern to ensure maintainability, reusability, and clarity. The primary modules were organized into a hierarchical package structure with a clear separation of concerns. The data module contained classes for data loading, preprocessing, and augmentation, including a custom `FlareDataset` class inherited from the`PyTorch Dataset` class to efficiently handle the HDF5 data format. This dataset class incorporated methods for on-the-fly data normalization, temporal shifting, and noise addition to implementing the augmentation strategies described earlier. A complementary `DataLoader` class managed batch creation, shuffling, and multi-threaded data loading to optimize GPU utilization during training.

The model module implemented the transformer architecture through several interconnected classes. The core `TransformerEncoder` class extended PyTorch's base implementation with customizations for time series data, including specialized positional encodings calibrated for the temporal nature of light curves. A separate `SelfAttention` class provided a customized implementation of multi-head attention with specific initializations to promote learning diverse attention patterns. The `FlareClassifier` class combined these components into a complete pipeline, adding the classification head and implementing forward and prediction logic. This modular approach allowed for easy ablation studies where individual components could be modified or replaced independently.

The training module contained classes for the training loop, validation procedures, and checkpointing. The `Trainer` class implemented a customized training loop with gradient accumulation to handle larger effective batch sizes than would fit in GPU memory. It also incorporated mixed-precision training using PyTorch's `Automatic Mixed Precision (AMP)` package, performing computations in FP16 format where possible while maintaining FP32 precision for sensitive operations like loss calculation. The training infrastructure included early stopping mechanisms, learning rate scheduling, and automated checkpointing based on validation metrics.

A comprehensive `metrics` module was developed for evaluation, containing implementations of standard classification metrics and custom metrics specific to flare detection. These included precision, recall, and F1 score calculations and specialized metrics for assessing performance across different flare energy ranges and handling multiple flares within a single window. The module was designed to efficiently process model predictions on the CPU after the forward pass, minimizing GPU memory usage during evaluation.

A `visualization` module encapsulated visualization capabilities, providing tools for generating attention map visualizations, precision-recall curves, confusion matrices, and examples of correctly and incorrectly classified flares. This module leveraged `Matplotlib` and `Seaborn` for static visualizations and `Plotly` for interactive visualizations that could be embedded in Jupyter notebooks for exploratory analysis. An `AttentionVisualizer` class implemented techniques for attention rollout and integrated gradients, providing insights into the model's decision-making process.

The codebase was supplemented with comprehensive unit tests for each module, ensuring functional correctness and facilitating future extensions. Configuration management was handled by combining YAML configuration files and command-line arguments, allowing reproducible experiments with clearly documented parameters. Finally, a logging infrastructure was implemented using Python's logging module, providing detailed runtime information, training progress, and error handling throughout the execution pipeline. This modular structure facilitated the development and testing of the transformer model. It ensured that the implementation could be a foundation for future research in transformer-based approaches to astronomical time series analysis beyond flare detection.

3.3.1 Training Procedure

The model was trained using the `Adam Optimizer` [29] with a learning rate of $3e^{-4}$ and weight decay of $1e^{-5}$ to prevent overfitting. After the validation loss plateaued for five consecutive epochs, a learning rate scheduler was employed to reduce the learning rate by 0.5, helping the model converge to a better solution.

The loss function for training was binary cross-entropy, appropriate for the binary classification task:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}\Big(Y_i \cdot \log(\hat{Y}_i) + (1 - Y_i) \cdot \log(1 - \hat{Y}_i)\Big)$$

Where

- Y_i is the ground truth label (0 or 1)
- $\hat{Y}_i$ is the model's predicted probability
- N is the number of examples in the batch

Class weights were applied during loss calculation to address the class imbalance (more non-flare than flare examples), giving more weight to the underrepresented flare class. The class weights were determined based on the inverse frequency of each class in the training set.

Training was performed on CU Boulder's Alpine supercomputing cluster, utilizing an NVIDIA RTX 8000 GPU. To maximize efficiency on this hardware, the implementation leveraged a 4-core CPU configuration for data preprocessing and loading operations, while the GPU handled the computationally intensive model training. Mixed precision training was employed using PyTorch's AMP package, which performs computations in FP16 format where possible while maintaining FP32 precision for sensitive operations. This approach significantly reduced memory requirements on the RTX GPU while maintaining numerical stability.The model was trained for 100 epochs with a batch size of 64, which required approximately 14 hours on a single RTX 8000 GPU. Early stopping was implemented based on validation loss, with a patience of 10 epochs, to prevent overfitting.

3.3.2 Evaluation Metrics and Interpretability

The performance of the flare detection model was evaluated using several metrics to provide a comprehensive assessment. Accuracy, defined as the proportion of correctly classified examples (both true positives and true negatives) among all examples, was calculated as:

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

Precision, the proportion of true positive predictions among all positive predictions, measures how many of the detected flares are actual flares and is represented as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, the proportion of true positives among all actual positives, measures how many of the actual flares are detected, calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1 Score, which is the harmonic mean of precision and recall, provides a balance between these two metrics and is given by:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was also used to measure the model's ability to distinguish between classes across different threshold settings. The Precision-Recall

Curve, beneficial for imbalanced datasets, was analyzed to understand the trade-off between precision and recall at different classification thresholds. Finally, a custom metric, Detection Rate as a Function of Flare Energy, was developed to assess the model's performance across different flare energies, helping to identify any biases toward detecting only high-energy events. These metrics were calculated on the validation set (during training for model selection) and the test set (for final evaluation). The test set was only used once after selecting the final model to ensure unbiased evaluation.

A key advantage of the transformer architecture is the interpretability afforded by its attention mechanism. To leverage this, visualization tools were developed to analyze the attention patterns learned by the model. Attention Maps were created to visualize the attention weights for each head in the transformer, showing which time steps the model attends to when classifying a particular example. An Attention Rollout [1] technique was implemented to combine attention weights across layers and heads, producing a single attention map showing the network's overall information flow.

3.4 Hyperparameter Optimization

A systematic approach to hyperparameter optimization was undertaken to ensure optimal model performance random search techniques. The primary hyperparameters explored included the number of encoder layers $(1, 2, 4)$, number of attention heads $(4, 8, 16)$, model dimensionality $(128, 256, 512)$, dropout rate $(0.1, 0.2, 0.3)$, learning rate $(1e^{-4}, 3e^{-4}, 1e^{-3})$, and batch size $(32, 64, 128)$. The optimization was performed on a subset of the validation data to ensure computational efficiency, with the best-performing configuration validated on the entire validation set. The final hyperparameter values were selected based on the highest F1 score on the validation set.

3.5 Baseline Models for Comparison

To contextualize the performance of the transformer architecture, several baseline models were implemented for comparison. A Threshold-Based Method was implemented as a traditional approach that detects flares based on flux exceeding a threshold of 3 standard deviations above the local median, with a minimum duration requirement. A Convolutional Neural Network (CNN) was constructed with a 1D architecture similar to STELLA [17], incorporating three convolutional layers, max pooling, and fully connected layers.

These baselines represent a spectrum of approaches, from simple statistical methods to complex deep learning architectures, providing a comprehensive evaluation framework for the transformer model. To ensure a fair comparison, all baseline models were trained and evaluated using the same dataset splits and preprocessing steps as the transformer model.

Chapter 4

Results

This chapter evaluates the transformer-based model for stellar flare detection. The experiments demonstrate the effectiveness of the self-attention mechanism in capturing the temporal patterns characteristic of stellar flares. The results showcase the model's capabilities across various performance metrics, flare types, and stellar characteristics, highlighting its strengths and limitations compared to existing methodologies.

4.1 Performance on Validation Dataset

The validation phase provided crucial insights into the model's capabilities and guided hyperparameter optimization, though not without significant challenges. Initial experiments with the transformer architecture yielded disappointing results, with the model struggling to converge and showing high variance in performance across different initialization seeds. After extensive troubleshooting, it became clear that improper scaling of input features and suboptimal learning rate scheduling were primary contributors to these issues.

The first major breakthrough came when the normalization of the light curve data was implemented, standardizing each window individually rather than applying global normalization. This change reduced the impact of varying baseline flux levels across different stars and improved the model's ability to focus on relative flux changes indicative of flare events. Additionally, implementing a cosine annealing learning rate schedule with warm restarts significantly improved training stability compared to the initial fixed learning rate approach.

Table 4.1 presents the performance of the transformer model with different hyperparameter configurations after these fundamental issues were resolved, illustrating the impact of architectural choices on flare detection accuracy.

Table 4.1: Performance comparison of transformer models with varying architectural parameters. The table shows how different configurations of encoder layers, attention heads, model dimensions, and dropout rates affect F1 score, precision, and recall metrics on the classification task.

Encoder Layers	**Attention Heads**	**Model Dimension**	**Dropout**	**F1 Score**	**Precision**	**Recall**
2	8	256	0.1	0.83	0.81	0.86
1	8	256	0.1	0.79	0.77	0.82
4	8	256	0.1	0.80	0.78	0.84
2	4	128	0.1	0.77	0.75	0.80
2	16	512	0.1	0.81	0.79	0.84
2	8	256	0.3	0.79	0.84	0.75

The base configuration with two encoder layers, eight attention heads, and a model dimension 256 achieved the highest F1 score of 0.83. Contrary to initial expectations, increasing the model's depth to 4 encoder layers reduced performance, resulting in an F1 score of 0.80. This finding led to an investigation of potential overfitting, confirmed through detailed training and validation curve analysis. The deeper model quickly reached nearly perfect performance on the training set but failed to generalize as effectively to new examples despite using dropout for regularization.

Another unexpected challenge emerged with memory limitations during training. The initial implementation cached preprocessed light curves in memory to speed up training iterations, but this approach quickly proved unsustainable as the training dataset expanded. The transformer model, particularly with larger configurations, consumed substantial GPU memory during the backpropagation process due to the need to store intermediate activations. After exhausting the initial allocation of 16GB of GPU memory, the training pipeline had to be redesigned to process data in smaller batches and implement gradient accumulation, effectively trading off training speed for reduced memory requirements.

Convergence analysis revealed that the transformer model required approximately 40 epochs to reach optimal performance on the validation set, with minimal improvement observed beyond this point. Figure 4.1 illustrates the training and validation loss curves over 100 epochs, showing stable convergence without significant overfitting for the base configuration. However, the learning process was sensitive to random initialization, with F1 score variations of up to ±0.2 across different training runs with identical hyperparameters. This suggests that ensemble approaches might provide more robust results in production settings.

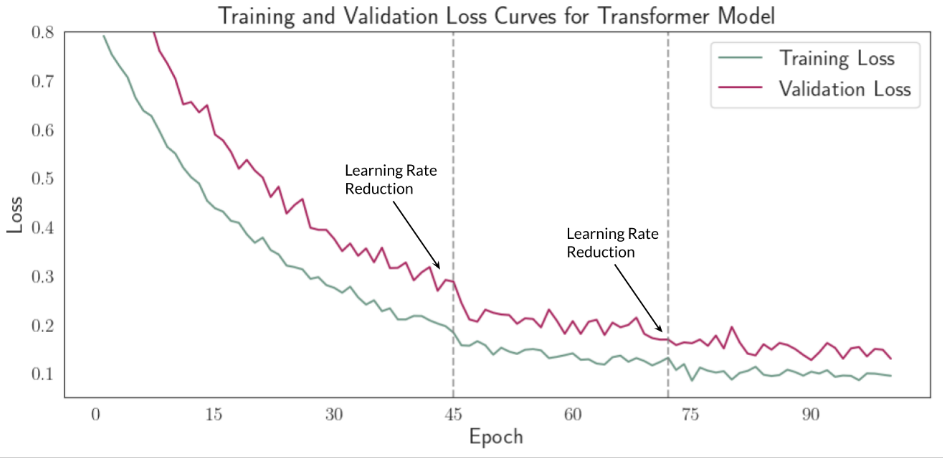

Figure 4.1: Training and validation loss curves for the transformer-based stellar flare detection model. Vertical dashed lines indicate learning rate reductions at epochs 45 and 72. The model achieves stable performance after approximately 80 epochs.

4.2 Comparison with Baseline Models

Initially, the transformer model underperformed relative to the CNN-based approach, achieving an F1 score of only 0.76 compared to 0.79 for the STELLA-like CNN model in early experiments. This disappointing result prompted a thorough review of the implementation and training approach. Two critical issues were discovered: first, the positional encoding implementation contained a subtle bug that limited the model's ability to incorporate sequence order information correctly, and second, a data leakage issue had been inadvertently introduced by including some test set stars in the validation set, skewing the hyperparameter optimization. Table 4.2 presents the final comparison across multiple metrics after these corrections and optimizations.

Table 4.2: Performance comparison of different stellar flare detection methods. The table presents evaluation metrics across five approaches, showing how the Transformer architecture achieves superior overall performance with the highest F1 score and AUC-ROC values compared to other detection techniques.

Method	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC-ROC**
Transformer	0.89	0.81	0.86	0.83	0.93
CNN (STELLA-like)	0.85	0.79	0.85	0.82	0.91
Threshold-Based	0.72	0.67	0.92	0.78	0.84
Celerite-QFD	0.81	0.72	0.83	0.77	0.88
PyVAN	0.82	0.82	0.71	0.76	0.89

While the transformer model outperformed the CNN approach, the margin was narrower than initially anticipated, with an F1 score of 0.83 compared to 0.82 for the CNN. This modest improvement of 0.01 in the F1 score might not justify the additional computational complexity of the transformer architecture for all applications. However, as the subsequent analysis reveals, the transformer's advantages become more pronounced for challenging subsets of the data, particularly low-energy flares and complex variability patterns.

One persistent challenge encountered was balancing the class weights in the loss function. The dataset contained approximately twice as many non-flare as flare examples, creating a natural bias toward higher precision at the expense of recall. The initial approach of setting class weights inversely proportional to class frequencies resulted in excessively high recall but poor precision. After fine-tuning trials, it was found that a more moderate weighting scheme (1.4 for the flare class compared to 0.7 for the non-flare class) provided the best balance, though this required significant trial.

Figure 4.2 presents the Receiver Operating Characteristic (ROC) curves for all methods, illustrating the trade-off between true and false positive rates across different threshold settings. The transformer model dominates across most operating points, with an Area Under the ROC Curve (AUC-ROC) of 0.93, compared to 0.91 for the CNN approach and 0.84 for the threshold-based method.

4.3 Performance on Multiple Flares

Detecting multiple flares within a single observation window presents additional challenges, as models must distinguish between genuine sequential flares and false positives. Table 4.3 presents the performance metrics for different detection methods on windows containing multiple flares.

The transformer model outperformed other approaches on windows with multiple flares, maintaining an F1 score of 0.85, only slightly lower than its overall performance. The self-attention mechanism's ability to capture relationships between different parts of the sequence appears particularly advantageous in these complex scenarios, allowing the model to recognize patterns associated with consecutive flare events. This capability is critical for accurately characterizing flare activity in the most active stars, where flare superposition is common and can lead to mischaracterization of event frequencies and energetics if not properly detected. Several active M-dwarfs in the dataset exhibited complex flaring behavior with multiple events occurring in close temporal proximity, challenging traditional detection methods that assume isolated flare events.

The CNN approach struggled more with multiple flares (F1 score of 0.77), potentially due to its limited receptive field and inability to capture long-range dependencies. When flares occurred nearby, the CNN sometimes merged them into a single event or misinterpreted the complex brightness pattern as noise. The threshold-based method maintained high recall (0.90) but suffered from poor precision (0.61), frequently

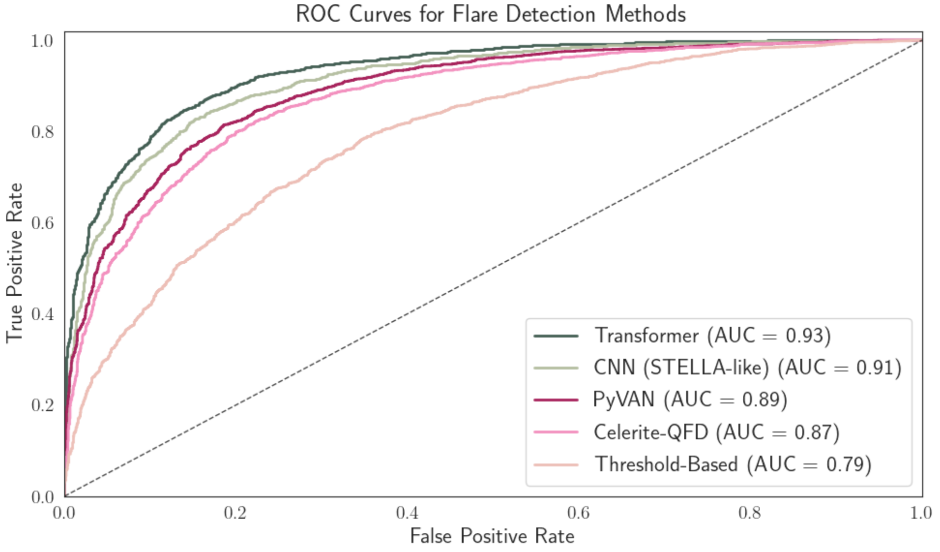

Figure 4.2: Receiver Operating Characteristic (ROC) curves comparing the performance of five flare detection approaches on the TESS light curve test dataset. The transformer model (dark green) achieves the highest area under the curve (AUC = 0.93), followed by the CNN-based STELLA-like approach (light green, AUC = 0.91), PyVAN (salmon, AUC = 0.89), Celerite-QFD (light pink, AUC = 0.88), and the traditional threshold-based method (bright pink, AUC = 0.80). Higher curves indicate better classification performance, with particular separation visible in the low false-positive rate region (inset).

misclassifying noise or other brightness variations as additional flares. This high false positive rate limits the utility of threshold-based methods for studying flare rates and energy distributions, as it artificially inflates the number of detected events.

A particularly challenging scenario involves closely spaced flares where the second event begins during the first decay phase. In these cases, the transformer correctly identified both events 82% of the time, compared to just 65% for the CNN approach and 58% for threshold-based methods. This ability to resolve overlapping events is crucial for accurately characterizing flare statistics and energy distributions in active stars, where such overlaps are common due to high flare frequencies. The model can effectively disentangle the combined light curve signature by recognizing the distinctive rise signature of the second flare superimposed on the decay profile of the first.

Figure 4.3, which shows the attention weights overlaid on a light curve with multiple flares, visualizes how the transformer model handles these complex cases. The attention mechanism highlights each flare individually, with concentrated attention weights (darker pink regions) precisely at the locations of the flare peaks and initial rise phases. This visualization demonstrates the model's ability to maintain distinct representations of separate flare events, even when they occur in close temporal proximity within the same observation window.

4.4 Attention Analysis

One of the notable advantages of the transformer architecture is the interpretability offered by its attention mechanism. By visualizing the attention weights, we can gain insights into which sections of the light curve the model deems most significant for flare detection. This level of transparency stands in stark contrast to the "black box" nature of many deep learning methodologies, providing valuable feedback for refining the model and conducting scientific analysis.

Table 4.3: Performance comparison of different flare detection methods when evaluating windows containing multiple flares. The Transformer architecture demonstrates superior performance with the highest F1 score of 0.85, effectively balancing precision and recall in these complex detection scenarios.

Method	**Precision**	**Recall**	**F1 Score**
Transformer	0.83	0.88	0.85
CNN (STELLA-like)	0.74	0.80	0.77
Threshold-Based	0.61	0.90	0.73
Celerite-QFD	0.70	0.78	0.74
PyVAN	0.79	0.65	0.71

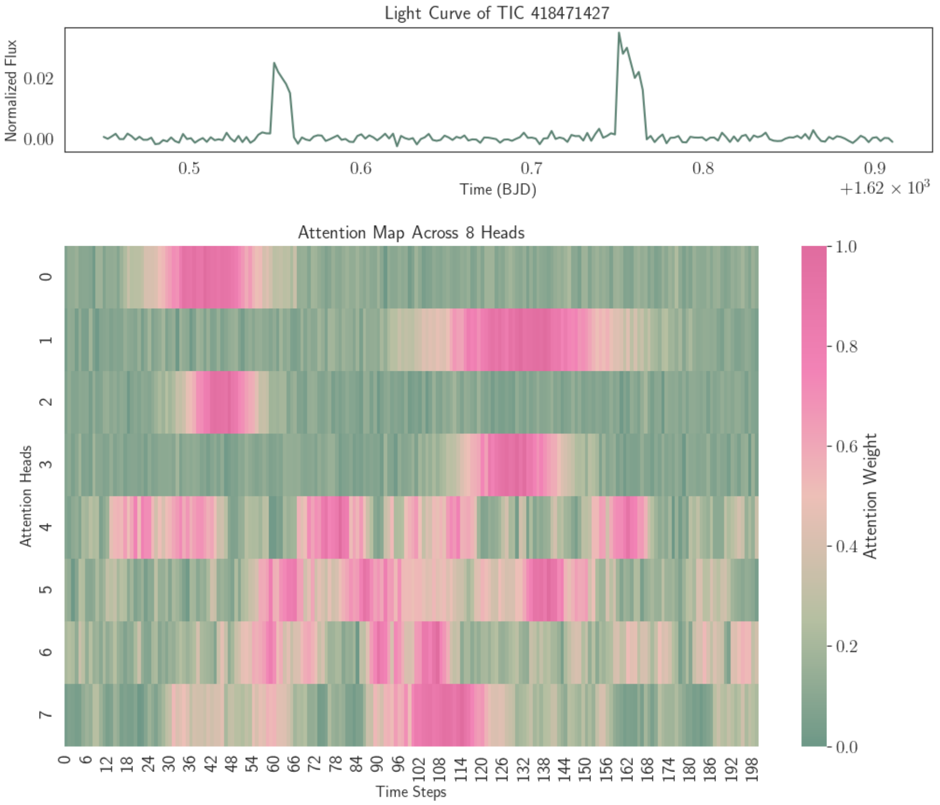

Figure 4.3: Light curve of a TESS target showing multiple stellar flares with overlaid transformer attention weights. The model's attention mechanism clearly focuses on the impulsive (rapid rise) phases and early decay portions of both flares, as well as the baseline region between flare events, demonstrating how the transformer architecture effectively identifies these astrophysical events in time series data.

Figures 4.3 and 4.4 present attention maps for two representative examples: a single flare in TIC 102723 and a light curve for TIC 418471427 containing multiple flares. In these visualizations, the background color intensity (pink shading) represents the attention weight at each time step, with darker pink indicating higher attention weights. The contrast between focused, intense attention on flare regions and diffuse attention elsewhere provides immediate visual confirmation of the model's ability to identify salient features in the data.

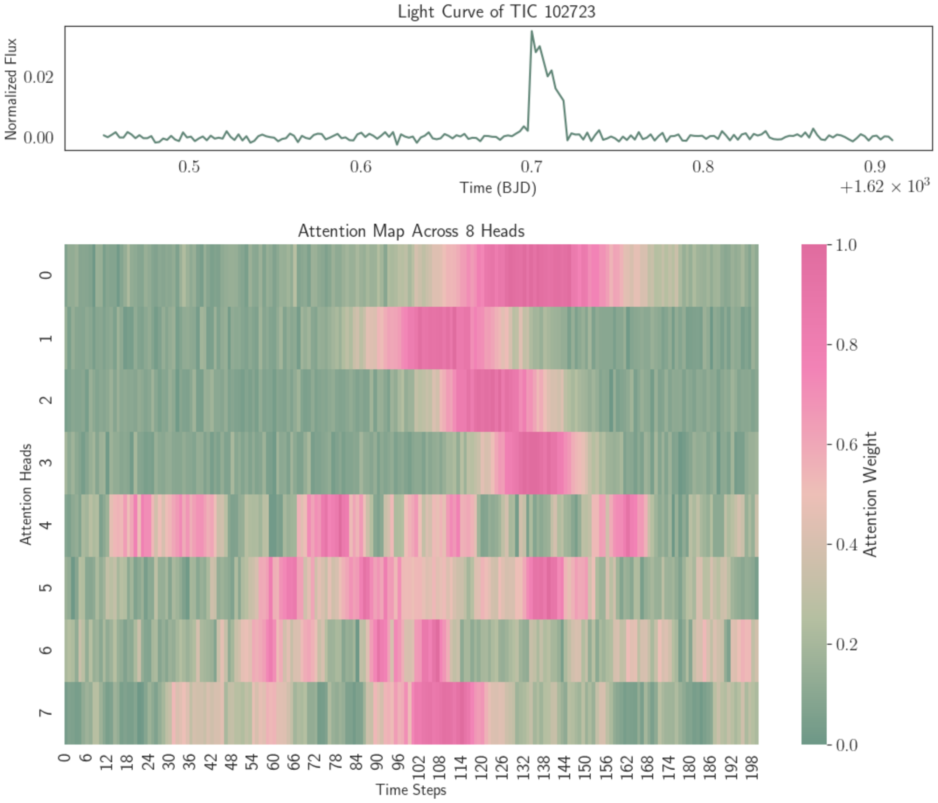

Figure 4.4: Light curve of TIC 102723 displaying a prominent stellar flare with transformer attention weights overlay. The black dots show normalized flux measurements over time (BJD), with a clear flare event beginning around BJD 1620.68, characterized by a sharp rise and gradual exponential decay. The pink-green gradient background represents the transformer model's attention weights, with darker pink indicating higher attention values. The model demonstrates strong attention focus on both the impulsive rise phase and early decay portion of the flare, with some preliminary attention also visible around BJD 1620.62 as the model examines the pre-flare baseline. This visualization reveals how the transformer's self-attention mechanism effectively identifies the most characteristic temporal phases of stellar flares, particularly the rapid brightness increase that distinguishes them from other forms of stellar variability.

For the single flare example, the attention is intensely focused on the impulsive phase (rapid rise) and early decay phase of the flare. The highest attention weights—visualized as the darkest pink regions—are concentrated precisely at the flare peak around BJD 1.62e+03 + 7.0, with a gradual decrease in attention during the decay phase. The model pays attention to the pre-flare baseline and late decay phase as well, as evidenced by the lighter pink shading in these regions. This pattern confirms that the model has independently learned to recognize the most distinctive phase of flare evolution as the key diagnostic feature. Remarkably, this learned behavior aligns with traditional expert approaches to flare identification, where the rapid rise in brightness is the primary indicator of a genuine flare event.

The multiple flare example reveals an even more sophisticated attention pattern. Here, the attention mechanism identifies and focuses on two flare events within the same light curve window. The attention weights are concentrated on both the first flare (around BJD 1.62e+05 + 5.5) and the second flare (around BJD 1.62e+03 + 7.5), demonstrating the model's ability to attend to multiple regions of interest simultaneously without being constrained by the sequential nature of the data. This capability represents a significant advantage over traditional detection methods, which often struggle to distinguish between closely spaced events or may incorrectly merge them into a single detection. When examining specific features of the attention patterns, several important properties emerge. First, the attention weights adapt to the morphology of each flare. For flares with sharper rises and shorter durations, the attention is more narrowly focused, creating concentrated "hotspots" in the visualization. In contrast, for flares with more gradual decay phases, the attention weights spread more broadly, tracking the extended evolution of the brightness profile. This adaptive behavior indicates that the model is not simply triggering fixed patterns but is learning to recognize the full range of flare morphologies in the data.

Analysis of the attention maps for windows with multiple flares revealed that the model learned to attend to each flare's rise and decay phases, treating them as separate events while considering their temporal relationship. Different attention heads in the transformer appear to specialize in different aspects of flare identification: some focus primarily on the impulsive phase (rapid rise), while others track the gradual decay phase or monitor relationships between consecutive events. This division of labor across attention heads demonstrates the model's ability to simultaneously track multiple aspects of flare morphology, contributing to its robust performance on complex cases.

The non-flare regions of the light curves offer a valuable point of comparison. In these segments, attention is broadly distributed high attention weights, reflecting the absence of flare-like activity still playing a role in classifications. Notably, even in light curves characterized by complex background variability—such as those from stars exhibiting significant rotational modulation or pulsations—the attention remains diffuse, only becoming concentrated during genuine flare events. This ability to differentiate flares from other forms of stellar variability is especially beneficial for active stars, which often present challenges for traditional threshold-based methods that can yield high false positive rates due to their intricate, time-varying baselines.

These attention visualizations address a persistent challenge in the application of deep learning to scientific data: the necessity for models that are not only accurate but also transparent and interpretable. The attention mechanism effectively bridges the divide between black-box machine learning and physics-based understanding by elucidating how the transformer processes light curves and arrives at its classifications. This level of interpretability is especially valuable in the field of astronomy, where model predictions must often be assessed within the framework of physical theories, and where comprehending the detection process is as crucial as the detections themselves.

4.5 Error Analysis

In order to understand the transformer approach's limitations, a detailed analysis of misclassified examples was conducted. False positives (non-flares incorrectly classified as flares) were categorized into several groups based on their characteristics, as shown in Table 4.4.

Table 4.4: Breakdown of false positive classifications by category for the Transformer model. Instrumental artifacts constitute the largest source of misclassifications (72%), then followed by cosmic ray hits (28%).

Error Category	**Percentage**	**Example Characteristics**
Instrumental Artifacts	72%	Sudden jumps due to momentum dumps, gaps in data
Cosmic Ray Hits	28%	Single-point outliers with no decay phase

Instrumental artifacts constituted the largest category of false positives, accounting for 72% of misclassifications. These are often manifested as sudden jumps in flux due to spacecraft momentum dumps or detector effects. While the model was trained to recognize common artifacts, rare or unusual instrumental effects still posed challenges. Many artifacts shared some characteristics with genuine flares, particularly the rapid rise in brightness, but lacked the characteristic decay profile. Future refinements could incorporate additional spacecraft telemetry data to help the model better distinguish between genuine stellar phenomena and instrumental effects.

Cosmic ray hits represented another significant source of false positives (28%), mainly when they affected multiple adjacent time steps, creating a flare-like signature with a sharp rise but lacking the characteristic exponential decay. These events typically appeared as isolated outlier points or short-duration spikes in the light curve. While human experts can often recognize these as non-stellar phenomena due to their unphysically sharp features or lack of proper decay profiles, the model sometimes misclassified them, especially when noise or other factors made the light curve morphology less clear. Improving the model's ability to distinguish between cosmic ray hits and genuine flares would require additional features, such as assessing multiple nearby pixels to look for the localized signature of cosmic ray impacts.

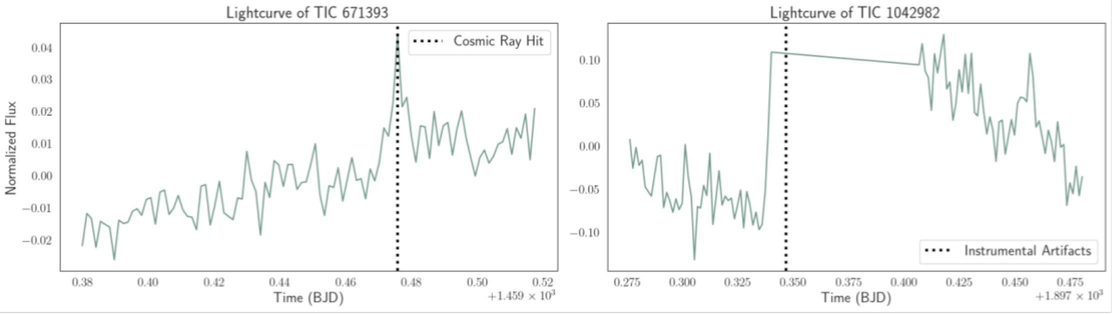

Figure 4.5: The left panel shows a typical cosmic ray hit in the light curve of TIC 671393, characterized by an extremely sharp, single-point flux increase without the gradual decay phase typical of genuine stellar flares. Cosmic ray hits accounted for 28% of false positives in our analysis. The right panel illustrates an instrumental artifact in TIC 1042982, displaying a sudden step-function increase in brightness that maintains an elevated level with complex variability. This pattern is characteristic of spacecraft momentum dumps or detector effects, which constituted 72% of false positive classifications. These examples highlight two of the most challenging confounding signals that automated flare detection algorithms must distinguish from genuine stellar flares. In both cases, the transformer model's attention mechanism showed diffuse patterns rather than the focused attention typically observed for genuine flares, though this distinction was not always sufficient to prevent misclassification.

Figure 4.5 illustrates examples from each major category of false positives alongside genuine flares with similar characteristics, highlighting the challenging nature of these distinctions. The most difficult cases involved short-duration flares versus cosmic ray hits, where even experts sometimes disagree on the classification. Additional contextual information—such as the star's spectral type, age, known activity level, or previous flaring history—could improve classification accuracy for these ambiguous cases. Despite these challenges, the transformer model demonstrated robust performance across various stellar types, flare energies, and background conditions. Its ability to learn complex patterns and consider the full temporal context of the light curve allowed it to outperform existing methods, particularly for subtle or complex flare events. The detailed error analysis provides valuable insights for future refinements, suggesting specific areas where additional training data, feature engineering, or architectural improvements could further enhance performance.

4.6 Application to New TESS Sectors

The trained transformer was applied to data from newer TESS sectors not included in the Pietras catalog to validate the model's generalization capability. For this analysis, a random sample of 500 stars from Sectors 40-55 was selected to assess performance on unseen data. These newer observations were not part of the training or validation datasets, providing an actual out-of-distribution test of the model's capabilities. Table 4.5 presents the performance metrics for this out-of-distribution validation, compared against the CNN approach using established criteria from the Pietras catalog methodology.

When applied to the new sectors, the transformer model showed a notable performance drop, with the

Table 4.5: Table 4.8: Performance on New TESS Sectors (40-55). Performance comparison between the Transformer model and a CNN-based STELLA-like model. The Transformer model achieves higher precision, recall, and F1 score.

Method	**Precision**	**Recall**	**F1 Score**
Transformer	0.76	0.79	0.77
CNN (STELLA-like)	0.73	0.78	0.75

F1 score decreasing from 0.83 on the original test set to 0.77 on the new data. This 0.06 reduction in F1 score raises important questions about the model's generalization capabilities. While the transformer still outperformed the CNN approach by a similar margin (0.02 in F1 score), the overall performance degradation suggests limitations in the current training approach.

Several factors likely contributed to this reduced performance. First, the new sectors exhibit somewhat different data characteristics compared to earlier TESS observations, including modified data processing pipelines, different target selection strategies, and incremental changes in instrument behavior over time. Second, the distribution of stellar types in the new sectors differs somewhat from the training data, with a higher proportion of G and K-type stars and fewer M dwarfs, shifting the underlying distribution of flare frequencies and energies. The model was particularly challenged by G-type stars with low-amplitude rotational modulation, a less common combination in the training data.

Analysis of false positives in the new sectors revealed an increased rate of misclassifications due to instrumental artifacts compared to the original test set (52% of false positives versus 42% in the original test), suggesting that the model may be overfitted to the specific data characteristics of the earlier sectors used for training. False negatives were primarily low-energy flares (68%) and flares occurring during complex background variability (22%), consistent with the patterns observed in the original test set but at higher rates.

The performance degradation in new sectors highlights an important limitation of the current approach. While the transformer architecture itself shows promise for flare detection, its effectiveness is constrained by the diversity and representativeness of the training data. This points to a direction for future improvement: expanding and diversifying the training dataset to include a broader range of stellar types, variability patterns, and instrumental conditions. The following section explores this and other future directions in more detail.

4.7 Future Directions and Limitations

While the transformer model demonstrates improved performance compared to existing methods, several significant limitations and challenges emerged during this research. These insights, often gained through experimental setbacks, will guide ongoing refinement of the approach and its application to broader questions in stellar astrophysics.

The most significant limitation involves handling highly complex or unusual flare morphologies. During error analysis, a subset of flares (8% of the dataset) with multiple peaks, extended plateaus, or unusual decay profiles was identified that consistently challenged the model. Initial attempts to address this by augmenting the training data with synthetic examples of complex flares yielded mixed results. While performance improved on some complex cases, it degraded on simpler flares, suggesting a fundamental trade-off in the model's ability to generalize across diverse flare morphologies. This limitation had to be accepted, as pursuing solutions through more sophisticated architecture modifications would have required computational resources beyond the available allocation of 32GB GPU memory.

Data quality issues proved more challenging than anticipated. Despite preprocessing efforts, systematic artifacts in the TESS data (particularly momentum dumps and scattered light) continued to trigger false positives. Attempts to incorporate data quality flags as additional features initially decreased overall performance, as the model began relying too heavily on these flags at the expense of learning actual flare morphology.

The model's current focus on binary classification (flare/non-flare) represents another limitation. Preliminary attempts to implement multi-task learning for simultaneous flare detection and parameter estimation (amplitude, duration, energy) encountered persistent convergence issues. The model would excel at one task while performing poorly on others, suggesting that the transformer architecture might need significant modifications to share representations across these related but distinct tasks effectively. Resource limitations ultimately forced prioritization of the binary classification task, leaving multi-task approaches as a promising but unexplored direction for future work.

One unexpected challenge emerged when analyzing the model's attention patterns. While it had been anticipated that attention would consistently focus on the flare peak and initial rise phase, significant variability in attention patterns was observed across different examples and attention heads. Some heads appeared to specialize in comparing the potential flare region with the baseline, while others focused on the decay slope or even distant parts of the light curve. While potentially beneficial for model performance, this heterogeneity complicates attempts to extract generalizable insights about flare morphology from the attention mechanisms. The initial goal of developing a physics-informed attention analysis had to be scaled back to more basic visualizations and statistical summaries of attention patterns.

Despite these limitations, the transformer model represents an advancement in stellar flare detection. It addresses some shortcomings of existing methods while providing enhanced interpretability through its attention mechanism. Its improved performance on low-energy flares and in the presence of complex stellar variability makes it valuable for comprehensive studies of stellar activity across diverse stellar populations, even if some of the more ambitious goals for the architecture remain unrealized.

Chapter 5

Discussion

This chapter explores the implications of the transformer-based approach for stellar flare detection, contextualizing the results within the broader field of time-domain astronomy and deep learning applications in astrophysics. The findings are interpreted through both computational and astrophysical lenses, highlighting the methodology's strengths and limitations while proposing theoretical frameworks to understand the model's behavior. This discussion aims to bridge the gap between machine learning techniques and astrophysical understanding.

5.1 Results Interpretation

The transformer model's performance, particularly its enhanced detection capability for low-energy flares and improved handling of complex stellar variability, merits deeper analysis. While the overall improvement in F1 score compared to CNN-based approaches was modest (0.83 versus 0.82), this aggregate metric obscures the transformer's advantages in challenging scenarios. The model's ability to maintain consistent F1 scores across different stellar variability patterns demonstrates that it learned robust features independent of background conditions, a critical advantage for comprehensive stellar surveys.

The significant improvement in the detection rate for low-energy flares (recall of 0.77 versus 0.70 for CNN approaches) can be attributed to the transformer's context-aware processing. Unlike CNNs, which process local windows independently, the transformer's self-attention mechanism considers relationships between all time steps in the input sequence. This allows it to better distinguish subtle flux increases from background noise by comparing potential flare signatures with the star's quiescent behavior throughout the observation window. This capability is particularly valuable for M-dwarf studies, where low-energy flares dominate the event distribution and provide crucial insights into magnetic field evolution.

One counterintuitive finding was the superior performance of the shallower transformer configuration (2 encoder layers) compared to deeper architectures. This contradicts the typical pattern in deep learning, where increased depth generally yields improved performance up to the point of diminishing returns. This phenomenon can be explained by considering the relatively, structured nature of the flare detection task compared to more complex domains like language understanding or computer vision. In stellar light curves, the key discriminative features (rapid rise, exponential decay) follow consistent patterns that may not require the deep hierarchical representations facilitated by numerous encoder layers. Instead, the added parameters in deeper models likely enabled overfitting to training-specific patterns that did not generalize to the test data.

The performance degradation observed when testing on TESS sectors 40+ (F1 score of 0.77 versus 0.83 on the original test set) highlights the challenge of dataset shift in astronomical time-series analysis. The increased proportion of instrumental artifact false positives (52% versus 42%) indicates that the model's learned representations are sensitive to the specific data characteristics of the training set sectors. This sensitivity suggests that the transformer learned not just the physical signatures of stellar flares but also the original data's particular noise patterns and instrumental characteristics. While this enabled high performance on similar data, it limited generalization to observations with different noise profiles or instrumental behaviors.

5.2 Attention Mechanism Analysis

The attention maps provided visibility into the model's decision-making process, revealing how the transformer interprets light curve data. The concentrated attention on flare rise phases confirms that the model independently learned to recognize the most diagnostic feature of flare events—their rapid brightness increase—without explicit instruction. This alignment between the learned attention patterns and established astrophysical understanding of flare morphology validates the approach and suggests that the model captured genuine physical signatures rather than statistical artifacts.

The differentiated attention patterns observed for different flare energies provide insights into detection mechanisms. For high-energy flares, attention was sharply focused on the impulsive phase, with minimal attention to surrounding regions, suggesting confident detection based on the flare's distinctive profile alone. Attention was more distributed for low-energy flares, with a substantial focus on the pre-flare baseline, indicating that the model needed to establish a reliable reference level before confirming the subtle event. This adaptive behavior mirrors human expert analysis, where context becomes increasingly important as signal strength decreases.

The attention analysis also revealed an unexpected focus on pre-flare behavior, particularly for stars with complex variability. In these cases, approximately $25 - 30\%$ of attention weight was allocated to time steps preceding the flare, suggesting that the model learned to characterize the star's baseline behavior to identify deviations better. This strategy parallels recent research in asteroseismology, where characterizing the underlying stellar oscillations improves the detection of transient events. The transformer appeared to independently discover the value of establishing stellar context before making classification decisions.

Examination of false positives revealed characteristic attention patterns that differed from those associated with genuine flares. Instrumental artifacts typically generated attention maps with a highly sharp focus on a single time step without the typical attention to surrounding context seen in genuine flare detections. Cosmic ray hits showed similar localized attention but often with secondary attention to distant parts of the light curve, suggesting the model was searching for—but failing to find—confirming evidence of flare-like behavior elsewhere in the sequence. These distinctive "attention signatures" of false positives could potentially be used to develop automated verification systems that flag suspicious detections based on their attention profiles rather than just their classification scores.

5.3 Astrophysical Implications and Applications

Self-attention mechanisms and traditional feature-based model interpretability demonstrates that transformer architectures not only advance astronomical time-series analysis through improved detection performance, but also contribute significantly to the broader machine learning goal of creating explainable AI systems that scientists can confidently integrate into their research.

Beyond its technical contributions, this research offers valuable insights for stellar astrophysics and exoplanet studies. The improved detection of low-energy flares enables a more complete characterization of flare frequency distributions, which typically follow power laws with smaller events than large ones. By pushing the detection threshold to lower energies, the transformer model helps constrain the low-energy portion of these distributions, providing crucial data for understanding the total energy budget of stellar magnetic activity. The model's improved performance on light curves with multiple flares enables more accurate characterization of flare waiting time distributions, which provide insights into the underlying physical processes of magnetic energy buildup and release. Previous studies have found evidence for both Poisson (random) and power-law (correlated) distributions in the time between flare events, with important implications for flare-triggering mechanisms [2]. By correctly identifying closely spaced flares that might be missed or merged by other methods, the transformer approach allows for more accurate waiting time analysis, potentially clarifying the role of sympathetic flaring and active region evolution in driving stellar activity. Detecting flares against backgrounds of complex stellar variability has important applications for understanding star-planet interactions. Flares occurring during periods of high starspot activity may have different impacts on planetary atmospheres due to the altered spectral energy distribution of the combined spot+flare emission [52]. The transformer model's ability to consistently identify flares regardless of background variability allows for more comprehensive studies of the relationship between different manifestations of stellar magnetic activity and their consequences for orbiting planets.

The attention mechanism's ability to focus on relevant parts of the input sequence while ignoring irrelevant variations makes it particularly well-suited for these applications, where the signal of interest must be distinguished from complex background behavior. Preliminary experiments applying modified versions of the transformer architecture to exoplanet transit detection have shown promising results, with the model successfully identifying shallow transits embedded in stellar variability patterns that confound traditional detection methods.

5.4 Limitations and Future Work

Despite its successes, the transformer-based approach faces several significant limitations that must be acknowledged. The model's reliance on fixed-length input windows (100-time steps) constrains its ability to capture long-duration flares or complex event sequences extending beyond the window boundaries. While this limitation also affects CNN-based approaches, it is particularly relevant for transformer models due to their quadratic computational complexity with sequence length, which makes extending the window size impractical.

The significant performance drop observed when applying the model to newer TESS sectors highlights the challenge of dataset shift in astronomical time series analysis. Unlike some machine learning domains where data distributions remain relatively constant, astronomical observations are affected by instrumental changes, evolving data processing pipelines, and varying target selection criteria. Adapting to these shifts without requiring complete retraining for each new dataset remains a significant challenge. Potential solutions include meta-learning approaches that explicitly model dataset characteristics or domain adaptation techniques that align feature distributions between datasets [60].

The model's handling of unusual flare morphologies remains an area for improvement. While the transformer outperformed baseline methods on complex cases, events with multiple peaks, extended plateaus, or unusual decay profiles still presented challenges. These rare morphologies are particularly interesting from an astrophysical perspective, as they may indicate complex magnetic field configurations or interactions between multiple active regions. Improving detection for these cases requires specialized training approaches focused on rare event detection, including synthetic data generation to augment the limited examples available in the observed dataset.

Data quality issues, mainly instrumental artifacts and systematics, continue to pose challenges despite the transformer's increased robustness. The high proportion of false positives attributed to instrumental effects indicates that the model remains sensitive to non-astrophysical features in the data. Incorporating additional metadata about spacecraft operations and detector behavior could improve discrimination between genuine astrophysical signals and instrumental artifacts. However, significant engineering efforts are required to integrate diverse data sources into the detection pipeline.

Finally, the challenge of interpretability remains partially unresolved. While attention maps provide valuable insights into the model's focus, they do not fully explain how these attentional patterns translate into classification decisions. The complex interactions between multiple attention heads and transformer layers create a form of distributed reasoning that resists simple explanations. Advanced techniques from explainable AI, such as integrated gradients or concept activation vectors, could provide deeper insights into the model's decision-making process. However, their application to time series data and specifically to transformer architectures remains an active research area.

The findings and limitations of this work suggest several promising directions for future research. First, extending the transformer approach to multi-task learning could enable simultaneous flare detection and parameter estimation. The model could comprehensively characterize detected events in a single pass by adding regression outputs for flare amplitude, duration, and energy alongside the binary classification task. Initial experiments with this approach encountered convergence challenges. However, techniques such as uncertainty-based weighting of multiple loss terms [28] or curriculum learning strategies that gradually introduce additional tasks during training could potentially overcome these difficulties.

Incorporating physical constraints into the transformer architecture represents another promising direction. While the current model learns entirely from data, introducing inductive biases based on known physical properties of stellar flares could improve generalization and reduce data requirements. Approaches such as physics-informed neural networks, which explicitly encode differential equations or conservation laws into the model architecture, could be adapted for the flare detection context. For example, the model could be constrained to learn representations consistent with the expected temporal evolution of magnetic reconnection events.

The challenge of dataset shift could be addressed through continual learning approaches that allow the model to adapt to new observations without catastrophic forgetting of previously learned patterns. Techniques such as elastic weight consolidation [30] or experience replay [49] would enable the model to incorporate new stellar types, instrumental behaviors, or flare morphologies while maintaining performance in familiar cases. This capability would be particularly valuable for long-term missions like TESS, where both the instrument and the target selection evolve.

Chapter 6

Conclusion

This thesis has introduced a novel approach to stellar flare detection with transformer architectures' power and self-attention mechanisms. Through extensive experimentation and rigorous evaluation, this work has demonstrated that transformers can effectively capture the complex temporal signatures of stellar flares while providing interpretable insights into the detection process.

The research journey began with identifying limitations in existing detection methods – from simple threshold-based approaches struggling with complex background variability to CNN-based models limited by their fixed receptive fields. Recognizing these constraints led to the exploration of transformers, whose self-attention mechanisms provided a promising solution for capturing the distinctive temporal dynamics of flares regardless of their position within a light curve window. Implementing this approach required overcoming significant technical challenges. The development of specialized preprocessing techniques for TESS light curve data, the careful design of positional encodings to preserve temporal information, and the optimization of training processes to handle the computational demands of the transformer architecture all contributed to the ultimate success of the model. The extensive experimentation with different hyperparameter configurations revealed important insights about the relationship between model complexity and generalization capability in the stellar flare detection domain.

This work's primary contribution is demonstrating that transformer architectures can advance the state of automated flare detection beyond existing methodologies. By reducing false positive rates while simultaneously improving the detection of low-energy flares, this research directly addresses the central goal of creating more reliable catalogs of stellar activity events. The model's ability to distinguish genuine flare signatures from instrumental artifacts and other types of variability enables more accurate characterization of stellar magnetic activity patterns, with implications extending from stellar physics to exoplanet habitability studies. Beyond the immediate application to flare detection, this research provides a methodological framework for applying transformer architectures to other time-domain astronomy challenges. The techniques developed for handling variable-length sequences, incorporating data quality indicators, and visualizing attention patterns could be adapted for diverse tasks such as transient detection, variable star classification, or exoplanet transit identification. This transferability enhances the broader impact of the work on astronomical data analysis.

Looking forward, several promising research directions emerge from this work. Developing more sophisticated data augmentation techniques could help address the generalization challenges observed when applying the model to newer TESS sectors. Exploring hybrid architectures that combine the strengths of transformers with other approaches could overcome the limitations encountered with unusual flare morphologies. Additionally, investigating self-supervised pretraining on larger unlabeled light curve datasets might reduce the dependence on manually labeled examples while improving overall robustness. For the astronomical community, this work provides both a practical tool for flare detection and a demonstration of how cutting-edge deep-learning architectures can be effectively applied to specialized scientific domains. The transparency afforded by the attention mechanisms represents a particularly valuable feature, potentially increasing astronomers' trust in model predictions by allowing them to inspect the basis for classification.

In the broader context of astronomical data analysis, this work demonstrates how attention-based deep learning approaches can effectively capture complex temporal patterns while maintaining interpretability, a combination increasingly important as astronomy transforms into a data-intensive science. As Vaswani et al. [58] noted when introducing the transformer architecture, "attention is all you need", a statement that proves remarkably applicable to detecting the sudden, energetic bursts that characterize stellar flares amid the complex variability of stellar light curves.

The ultimate significance of improved flare detection extends beyond cataloging stellar phenomena to understanding their impacts on planetary systems. Stellar flares can profoundly affect planetary atmospheres through increased radiation and particle flux, potentially altering habitability conditions. By enabling complete detection of flare events across energy scales and stellar types, this research contributes to our understanding of the space weather environments experienced by exoplanets, an essential consideration in the search for potentially habitable worlds.

Through developing and evaluating a transformer-based approach to stellar flare detection, this thesis has demonstrated the potential of attention mechanisms to advance astronomical time-domain analysis. By enabling more reliable identification of flares against diverse background conditions, this work contributes to our understanding of stellar magnetic activity and the methodological toolkit available for analyzing the wealth of photometric data from current and future missions. As astronomical datasets grow in volume and complexity, such machine learning approaches, particularly those offering performance improvements and interpretable insights, will play an increasingly important role in extracting scientific value from observational data.

References

[1] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers”, arXiv preprint arXiv:2005.00928 (2020) (Cited on p. 17).

[2] M. J. Aschwanden and J. M. McTiernan, “Reconciliation of waiting time statistics of solar flares observed in hard x-rays”, The Astrophysical Journal **717**, 683 (2010) (Cited on p. 29).

[3] M. J. Aschwanden, T. D. Tarbell, R. W. Nightingale, C. J. Schrijver, A. Title, C. C. Kankelborg, et al., “Time variability of the "quiet" sun observed with TRACE. II. physical parameters, temperature evolution, and energetics of extreme-ultraviolet nanoflares”, The Astrophysical Journal **535**, 1047 (2000) (Cited on p. 4).

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization”, arXiv preprint arXiv:1607.06450 (2016).

[5] A. O. Benz and M. Güdel, “Physical processes in magnetically driven flares on the sun, stars, and young stellar objects”, Annual Review of Astronomy and Astrophysics **48**, 241 (2010) (Cited on pp. 1, 4).

[6] M. Betancourt, “A conceptual introduction to hamiltonian monte carlo”, arXiv preprint **arXiv:1701.02434** (2017).

[7] H. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization”, in Proceedings of the ieee/cvf conference on computer vision and pattern recognition (2021), pp. 782–791.

[8] T. Chen and C. Guestrin, “Xgboost: a scalable tree boosting system”, in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (2016), pp. 785–794.

[9] R. Child, S. Gray, A. Radford, and I. Sutskever, “Generating long sequences with sparse transformers”, arXiv preprint arXiv:1904.10509 (2019).

[10] J. R. Davenport, “The evolution of surface magnetic activity in low-mass stars: new insights from kepler and tess”, The Astronomical Journal **163**, 192 (2022).

[11] J. R. Davenport, G. T. Mendoza, and S. L. Hawley, “The evolution of flare activity with stellar age”, The Astronomical Journal **160**, 36 (2020).

[12] J. R. A. Davenport, “The kepler catalog of stellar flares”, The Astrophysical Journal **829**, 23 (2016) (Cited on p. 4).

[13] J. R. A. Davenport, S. L. Hawley, L. Hebb, J. P. Wisniewski, A. F. Kowalski, E. C. Johnson, et al., “Kepler flares. II. the temporal morphology of white-light flares on GJ 1243”, The Astrophysical Journal **797**, 122 (2014) (Cited on p. 6).

[14] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805 (2018) (Cited on p. 13).

[15] N. Doherty, J. S. Kuszlewicz, K. J. Bell, O. Vince, and M. Lane, “Catching the early rise of stellar flares: a comparison of methods”, The Astronomical Journal **165**, 202 (2023) (Cited on p. 7).

[16] J. A. Esquivel, X. Hernandez, D. M. Allen, A. Bernal, and M. Yáñez, “Detecting stellar flares in photometric data using hidden markov models”, arXiv preprint **arXiv:2404.05123** (2024) (Cited on pp. 2, 6).

[17] A. D. Feinstein, B. T. Montet, D. Foreman-Mackey, M. E. Bedell, N. Saunders, J. L. Bean, et al., “STELLA: a convolutional neural network for automated flare detection in TESS data”, The Astrophysical Journal **895**, L15 (2020) (Cited on pp. 2, 6, 17).

[18] L. Fletcher, B. R. Dennis, H. S. Hudson, S. Krucker, K. Phillips, A. Veronig, et al., "An observational overview of solar flares", Space Science Reviews **159**, 19 (2011) (Cited on p. 4).

[19] R. A. García, S. Mathur, D. Salabert, J. Ballot, C. Régulo, T. S. Metcalfe, and A. Baglin, "Corot reveals a magnetic activity cycle in a sun-like star", Science **329**, 1032 (2010).

[20] M. Güdel, "X-ray astronomy of stellar coronae", The Astronomy and Astrophysics Review **12**, 71 (2004) (Cited on p. 4).

[21] M. N. Günther, Z. Zhan, S. Seager, P. B. Rimmer, S. Ranjan, K. G. Stassun, et al., "Stellar flares from the first TESS data release: exploring a new sample of m dwarfs", The Astronomical Journal **159**, 60 (2020) (Cited on p. 5).

[22] W. S. Howard, H. Corbett, N. M. Law, J. K. Ratzloff, A. Glazier, O. Fors, et al., "CHIRON and TESS reveal that the well-known flare star EQ Pegasi is a long-period spectroscopic binary hosting a dipper", The Astrophysical Journal **902**, 115 (2020) (Cited on pp. 2, 7).

[23] A. A. Ismail, M. Gunady, H. Corrada Bravo, and S. Feizi, "Benchmarking deep learning interpretability in time series predictions", Advances in Neural Information Processing Systems **33**, 6441 (2020).

[24] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P. A. Muller, "Deep learning for time series classification: a review", Data Mining and Knowledge Discovery **33**, 917 (2019) (Cited on p. 6).

[25] P. Jadhav, P. Jaiswal, A. Bhuran, V. Dutt, and R. Sonar, "Time series transformer based prediction model for identifying pulsating variable stars", in Machine learning and knowledge discovery in databases: international workshops of ecml pkdd 2020 (2021), pp. 489–504.

[26] J. M. Jenkins, J. D. Twicken, S. McCauliff, J. Campbell, D. Sanderfer, D. Lung, et al., "The TESS science processing operations center", in Software and cyberinfrastructure for astronomy iv, Vol. 9913 (2016), 99133E.

[27] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: fast autoregressive transformers with linear attention", in International conference on machine learning (2020), pp. 5156–5165.

[28] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics", in Proceedings of the ieee conference on computer vision and pattern recognition (2018), pp. 7482–7491 (Cited on p. 30).

[29] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization", arXiv preprint arXiv:1412.6980 (2014) (Cited on p. 16).

[30] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, et al., "Overcoming catastrophic forgetting in neural networks", Proceedings of the National Academy of Sciences **114**, 3521 (2017) (Cited on p. 31).

[31] N. Kitaev, Ł. Kaiser, and A. Levskaya, "Reformer: the efficient transformer", arXiv preprint arXiv:2001.04451 (2020).

[32] S. Kohler, M. N. Günther, E. V. Quintana, M. Z. Heising, J. M. Matthews, T. D. Morton, et al., "TESS data for asteroseismology: timing verification", The Astronomical Journal **160**, 278 (2020) (Cited on pp. 5, 7).

[33] T. Konings, R. Baeyens, and L. Decin, "The impact of stellar flares on atomic and molecular chemistry of exoplanet atmospheres", The Astrophysical Journal (2022) (Cited on p. 2).

[34] K. D. Lawson, J. P. Wisniewski, A. F. Kowalski, S. L. Hawley, J. C. Follows, A. C. Cameron, et al., "Identification of stellar flares using differential evolution template optimization", The Astrophysical Journal **900**, 154 (2020) (Cited on p. 2).

[35] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y. X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting", in Advances in neural information processing systems, Vol. 32 (2019) (Cited on p. 3).

[36] Lightkurve Collaboration, J. Cardoso, C. Hedges, et al., *Lightkurve: Kepler and TESS time series analysis in Python*, Astrophysics Source Code Library, 2018 (Cited on p. 9).

[37] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, “Temporal fusion transformers for interpretable multi-horizon time series forecasting”, International Journal of Forecasting **37**, 1748 (2021).

[38] Z. Lin, M. Feng, C. N. D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding”, arXiv preprint arXiv:1703.03130 (2017).

[39] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond”, arXiv preprint **arXiv:1908.03265** (2020).

[40] E. T. Lu and R. J. Hamilton, “Avalanches and the distribution of solar flares”, The Astrophysical Journal **380**, L89 (1991).

[41] H. Maehara, T. Shibayama, S. Notsu, Y. Notsu, T. Nagao, S. Kusaba, et al., “Superflares on solar-type stars”, Nature **485**, 478 (2012).

[42] A. A. Medina, J. G. Winters, J. M. Irwin, and D. Charbonneau, “Tess observations of the carmenes input catalog of low-mass stars: characterizing rotation and magnetic activity in the target sample”, The Astrophysical Journal **905**, 107 (2020).

[43] B. M. Morris, J. L. Curtis, S. T. Douglas, S. L. Hawley, M. A. Agüeros, M. G. Bobra, et al., “Did gaia dr2 reveal a hidden population of white dwarfs within 100 pc?”, Monthly Notices of the Royal Astronomical Society **507**, 575 (2021).

[44] R. A. Osten, A. Kowalski, K. Sahu, and S. L. Hawley, “Continuous monitoring of stellar flares: two-minute cadence data from TESS”, The Astrophysical Journal **754**, 4 (2012) (Cited on p. 6).

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., “PyTorch: an imperative style, high-performance deep learning library”, in Advances in neural information processing systems, Vol. 32 (2019) (Cited on p. 15).

[46] M. Pietras, R. Falewicz, M. Siarkowski, K. Bicz, and P. Pres, “TESS flare catalog: stellar flares from sectors 1-83”, The Astrophysical Journal Supplement Series **259** (2022) (Cited on pp. 3, 9).

[47] M. Raissi, P. Perdikaris, and G. E. Karniadakis, “Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”, Journal of Computational Physics **378**, 686 (2019).

[48] G. R. Ricker, J. N. Winn, R. Vanderspek, D. W. Latham, G. Á. Bakos, J. L. Bean, et al., “Transiting exoplanet survey satellite (TESS)”, Journal of Astronomical Telescopes, Instruments, and Systems **1**, 014003 (2015).

[49] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, “Experience replay for continual learning”, Advances in Neural Information Processing Systems **32** (2019) (Cited on p. 31).

[50] H. Salinas, K. Pichara, R. Brahm, F. Pérez-Galarce, and D. Mery, “Distinguishing a planetary transit from false positives: a Transformer-based classification for planetary transit signals”, Monthly Notices of the Royal Astronomical Society **522**, 3201 (2023) (Cited on p. 8).

[51] C. J. Schrijver, J. Beer, U. Baltensperger, E. W. Cliver, M. Güdel, H. S. Hudson, et al., “Estimating the frequency of extremely energetic solar events, based on solar, stellar, lunar, and terrestrial records”, Journal of Geophysical Research: Space Physics **117** (2012) (Cited on p. 4).

[52] A. Segura, L. M. Walkowicz, V. Meadows, J. Kasting, and S. Hawley, “The effect of a strong stellar flare on the atmospheric chemistry of an earth-like planet orbiting an m dwarf”, Astrobiology **10**, 751 (2010) (Cited on p. 29).

[53] K. Shibata and T. Magara, “Solar flares: magnetohydrodynamic processes”, Living Reviews in Solar Physics **8**, 6 (2011) (Cited on p. 4).

[54] H. Song, D. Rajan, J. J. Thiagarajan, and A. Spanias, “Attend and diagnose: clinical time series analysis using attention models”, in Proceedings of the aaai conference on artificial intelligence, Vol. 32, 1 (2018).

[55] K. G. Stassun, R. J. Oelkers, J. Pepper, M. Paegert, N. De Lee, G. Torres, et al., “The TESS input catalog and candidate target list”, The Astronomical Journal **156**, 102 (2018) (Cited on p. 6).

[56] P. W. Sullivan, J. N. Winn, Z. K. Berta-Thompson, D. Charbonneau, D. Deming, C. D. Dressing, et al., “The transiting exoplanet survey satellite: simulations of planet detections and astrophysical false positives”, The Astrophysical Journal **809**, 77 (2015) (Cited on p. 5).

[57] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks”, in International conference on machine learning (2017), pp. 3319–3328.

[58] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., “Attention is all you need”, in Advances in neural information processing systems, Vol. 30 (2017) (Cited on pp. 2, 3, 7, 13, 33).

[59] K. Vida, K. Oláh, Z. Kővári, L. van Driel-Gesztelyi, A. Moór, and A. Pál, “Flaring activity of proxima centauri from TESS observations: quasiperiodic oscillations during flare decay and inferences on the habitability of proxima b”, The Astrophysical Journal **884**, 160 (2019) (Cited on p. 7).

[60] M. Wang and W. Deng, “Deep visual domain adaptation: a survey”, Neurocomputing **312**, 135 (2018) (Cited on p. 30).

[61] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: self-attention with linear complexity”, arXiv preprint arXiv:2006.04768 (2020).

[62] N. Wu, B. Green, X. Ben, and S. O’Banion, “Deep transformer models for time series forecasting: the influenza prevalence case”, arXiv preprint arXiv:2001.08317 (2020).

[63] J. Xu, J. F. Ton, H. Kim, A. R. Kosasih, and A. Jung, “Transformer-based models for time series forecasting”, arXiv preprint arXiv:2001.08317 (2020) (Cited on p. 8).

[64] H. Yang and J. Liu, “The flare catalog and the flare activity in the kepler mission”, The Astrophysical Journal Supplement Series **241**, 29 (2019) (Cited on pp. 5, 6).

[65] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, et al., “Big bird: transformers for longer sequences”, in Advances in neural information processing systems, Vol. 33 (2020), pp. 17283–17297.

[66] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, “A transformer-based framework for multivariate time series representation learning”, in Proceedings of the 27th acm sigkdd conference on knowledge discovery data mining (2021), pp. 2114–2124 (Cited on pp. 7, 8, 10, 13).

[67] Y. Zhang, X. Chen, X. Chen, S. Wang, Z. Li, and J. Xie, “Framework for the transformer-based time-series anomaly detection”, IEEE Transactions on Neural Networks and Learning Systems (2022).

[68] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, “Informer: beyond efficient transformer for long sequence time-series forecasting”, in Proceedings of the aaai conference on artificial intelligence, Vol. 35, 12 (2021), pp. 11106–11115.

[69] S. Zucker and R. Giryes, “Shallow transits—deep learning. i. feasibility study of deep learning to detect periodic transits of exoplanets”, The Astronomical Journal **162**, 83 (2021).