Improving Short-term Solar Proton Event Forecasting with Electron-Based Inputs and Transformer Architectures

Aatiya Ali[1], Viacheslav Sadykov[1]
Berkay Aydin[1], Benedict Mervyn[1]
[1]Georgia State University

Solar proton events (SPEs), a solar energetic particle (SEP) event subclass, are defined as increases of ≥10 MeV proton fluxes. While rare, SPEs can elevate space radiation levels, creating a hostile environment for humans and technical systems. Given their destructive potential, developing reliable, operational prediction models is a vital frontier in space safety.

01 Motivation

Reliable operational prediction requires balancing lead time with historical context. This study evaluates a **Transformer** model against the **Sliding Window Multivariate Time Series Forest (Slim-TSF[1]**; an ensemble-based module from GSU's Data Mining Lab**)**. By testing these models, we analyze how attention-based mechanisms and sliding-window ensembles handle the data imbalance and temporal dependencies critical to forecasting rare SEP surges.

02 The Role of Relativistic Electrons

This work utilizes a longitudinal dataset spanning 1995–2017. This integrates **high-energy relativistic electron fluxes from ACE/EPAM (0.045–0.312 MeV)** and **SOHO/EPHIN (0.387–1.147 MeV)** as precursors for SPE onset at geostationary orbit (GEO). This is physically motivated by two phenomena:

- **Velocity Dispersion**: Propagating near the speed of light, relativistic electrons offer a 30-to-120 minute early-warning buffer, allowing models to identify proton surges before manifesting at GEO.
- **Magnetic Connectivity**: Increases in electron fluxes serve as tracers of the particles along the Parker Spiral field lines, confirming a direct path between the solar eruption and an impending geoeffective SPE.

We also include magnetospheric relativistic electron fluxes measured by GOES (0.6 - >4 MeV). The forecast target is structured from GOES proton data where an SPE is defined by the onset of ≥10 MeV proton flux exceeding the operational 10 pfu threshold. All other conditions are categorized as Non-events.

03 Model Architectures

To transform continuous flux data into a classification task, we employ the Slim-TSF ensemble which **segments time-series data into discrete training and prediction intervals.** In contrast, the Transformer utilizes **self-attention** to weigh temporal dependencies across the entire sequence simultaneously. Both approaches are governed by the lookback (historical context for feature extraction), and lead times (prediction horizons). This allows an evaluation of how effectively each architecture identifies impending SEP surges before they manifest at GEO.

Feature	Transformer Classifier (Deep Learning)	Slim-TSF (Ensemble)
Input handling	Processes 3D temporal windows directly, structured as [Batch Size, Time Steps, Features]	Flattens temporal data into 2D, structured as [Batch Size, Flattened Features]
Temporal awareness	Inherently sequential	Statistically driven; may overfit to specific points rather than overall trend
Core logic	Self-Attention learns global dependencies	Decision Trees find non-linear thresholds
Imbalance handling	Uses probability thresholding to adjust sensitivity	Uses 'balanced' weights to penalize minority class errors during training
Feature Selection	ANOVA-based selection on temporal means	Time-series mean and σ used to reduce false signals
Optimization	GridSearch for dropout, batch size, and epochs.	GridSearch for tree depth, estimators, and class weights

Table 1: Structural comparison of model architectures. By limiting optimization to key parameters, we evaluate the inherent capabilities of each framework for SPE prediction.

04 Preliminary Results

To evaluate model robustness and predictive stability, we conducted 40 independent simulations for each architecture (see Table 1). The forecasting framework utilized two **observation windows** (**24 and 72 hours**) to assess short- and long-term temporal dependencies across **four lead-times (15 min, 30 min, and 1 hr.)** Preliminary benchmarking results across these configurations follow.

Feature Selection

Both models identify the EPHIN 1.080 MeV and 3.854 MeV electron channels as the most critical features across all horizons.

- **Transformer**: Feature importances are significantly larger toward the 15-min lead time. As lead times increase, feature predictive weights drop sharply. The model relies heavily on the the EPHIN channels and the EPAM 0.045–0.062 MeV, 0.062–0.103 MeV, and 0.103–0.175 MeV electron bands.
- **Slim-TSF**: Maintains a more balanced importance profile, distributing weight evenly across lead times. Interestingly, it places high importance on GOES 0.6 - 4 MeV and >4 MeV electron fluxes

Comparative feature importance is illustrated in Figure 1, highlighting the relative weight assigned to these energy channels. Note differences in x-axis scales arising from the unique internal logic and scoring metrics used by each model to evaluate feature contribution.

Overall Performance

Benchmarking results across lead times reveal a clear performance hierarchy favoring ensemble-based architectures for short-term SPE forecasting. As illustrated by the TSS, HSS, and F1 scores in Figure 2, **Slim-TSF consistently maintains robust predictive power,** achieving scores between 0.6 and 0.75 on average.

In contrast, the Transformer performs is significantly weaker, pointing to fundamental difficulties in identifying complex temporal patterns required for accurate SPE forecasting.

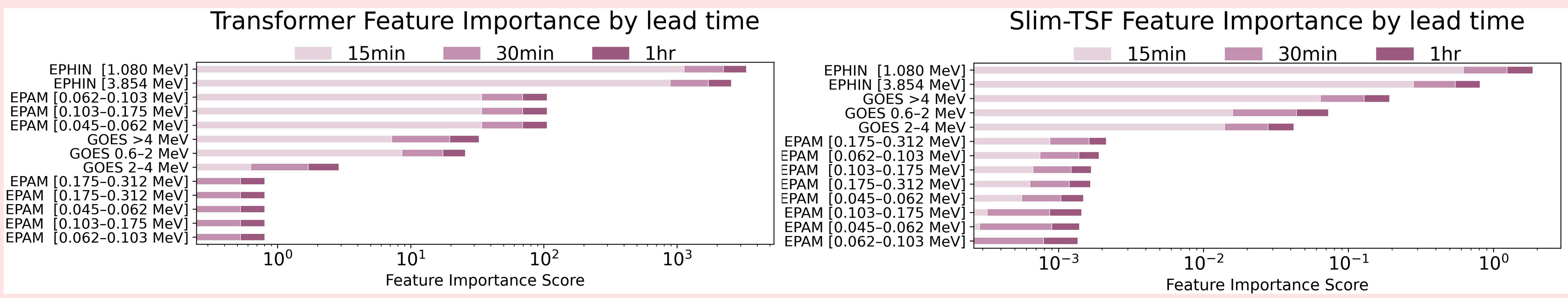

Figure 1: Normalized feature importance for all input energy channels. Scores are shown for the Transformer and SLIM-TSF models evaluated with lead times 15 min, 30 min, and 1 hr.

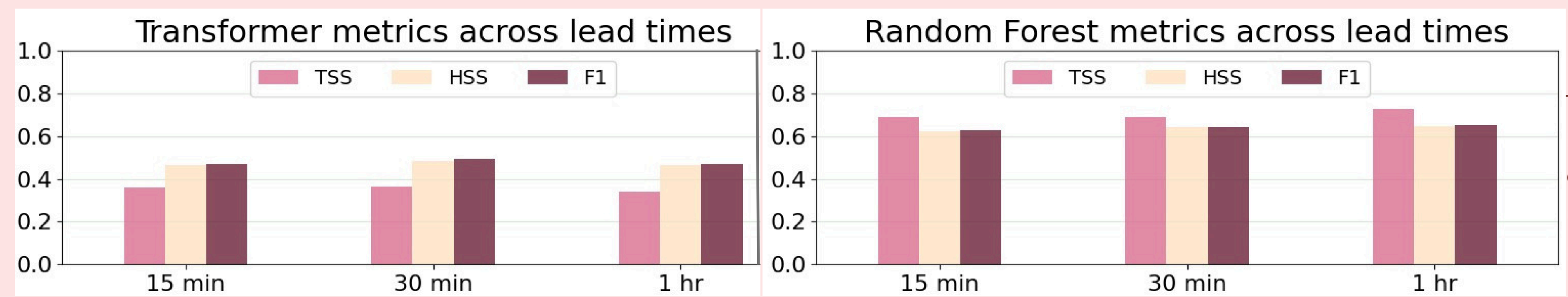

Figure 2: Model performance metrics across different lead times. Notably, metrics showed negligible variance across different observation windows (24h vs. 72h) and are therefore not shown here.

Overall Performance

Figure 3 clarifies this performance gap, where Slim-TSF demonstrates significantly higher operational reliability, **correctly identifying ~72% of actual SPEs** on average. In contrast, the **Transformer misses ~60% of events while correctly capturing only 40.1%** of hazardous SEP surges.

While the Transformer achieves near-perfect accuracy for non-events, this primarily reflects the dataset's extreme class imbalance, struggling to navigate rare-event thresholds essential for space safety.

Figure 4 illustrates how Slim-TSF significantly outperforms established benchmarks in short-term SEP forecasting. While Posner (2007)[2] relies on relativistic electron intensities within a physics-based statistical matrix and Torres et al. (2025)[3] use a neural network with time-series electron inputs, Slim-TSF achieves superior classification accuracy across both 30 min and 1 hr lead times. Specifically, **Slim-TSF yields an F1 score of 0.69 at 30 min and 0.58 at 60 minute lead times, exceeding the respective Torres (0.30, 0.16) and Posner (0.56, 0.31) baselines.**

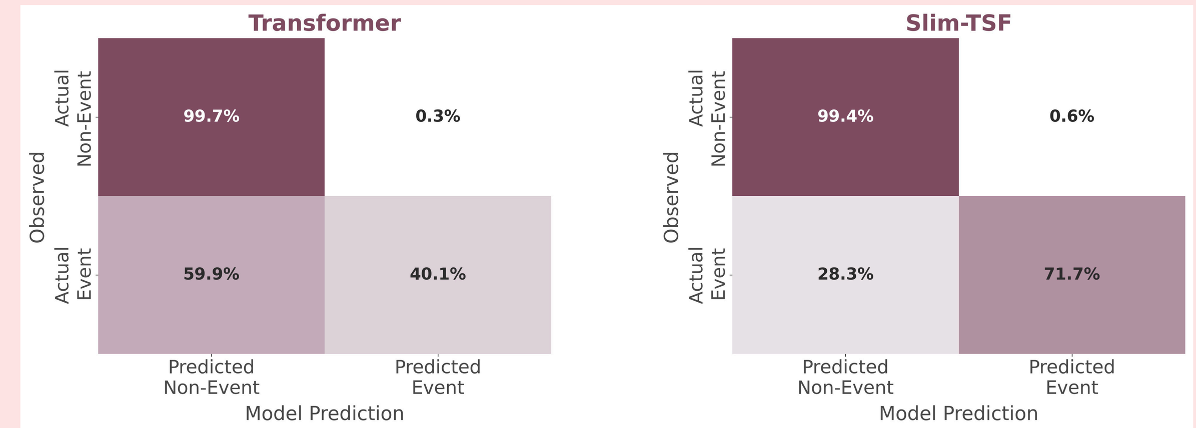

Figure 3: Confusion matrices showing the distribution of predictive outcomes. The dataset exhibits significant class imbalance, averaging ~25 positive cases (SPEs) contrasted with 2,000 negative cases (non-SPEs) per experimental run.

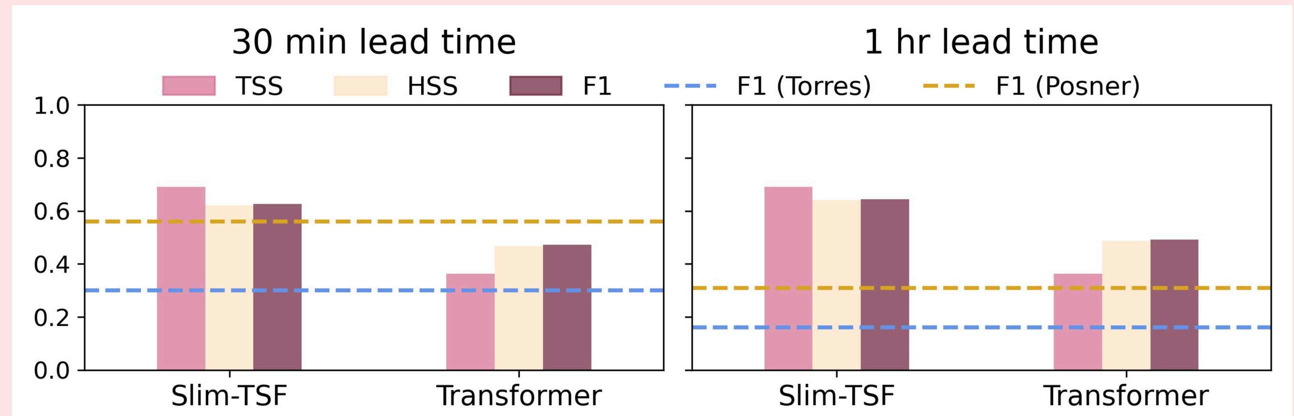

Figure 4: Comparing Slim-TSF and Transformer performances against Posner 2007 and Torres et al. 2025 benchmarks for 30 min and 1 hr lead times. While reference works use different input and architectures, the comparison highlights Slim-TSF's ability to maximize predictive accuracy for SEP events..

Key Takeaways

- Ensemble-based methods like Slim-TSF are more effective at integrating multi-instrument data from EPHIN, GOES, and EPAM sensors. **Slim-TSF provides a 25-30% improvement in TSS over the Transformer** at the critical 1 hour lead time.
- Although **both models identify the EPHIN 1.080 MeV and 3.854 MeV electron channels as primary precursors**, **Slim-TSF maintains a more balanced and stable feature importance distribution**. The Transformer exhibits more variance, indicating less reliable weight attribution across independent runs.
- Slim-TSF shows a robust increase in performance (peaking at TSS ~0.75) as lead times approach 1 hr. In contrast, the **Transformer appears to lack the temporal resilience** necessary for consistent event detection.
- Results suggest that the **Transformer's attention mechanism may be overly sensitive to localized noise**, causing it to struggle with the rare-event thresholds essential for space weather forecasting.
- A **negligible variance between 24 hr and 72 hr observation windows** suggest that predictive signatures are concentrated in the 24 hours preceding an SPE. This allows for significantly reduced data storage and computational overhead in operational environments without sacrificing predictive accuracy.
- **Slim-TSF achieves F1 scores of 0.69 and 0.58 for 30 and 60 minute leads; nearly doubling baseline results** and demonstrating an enhanced ability to capture non-linear SEP relationships.
- By utilizing a sliding-window ensemble to segment time-series data, Slim-TSF captures aggregate temporal trends, prioritizing long-term trend signatures to ensure a more stable decision boundary for operational forecasting.

References & Acknowledgements

[1] Ji, A. et al. (2024). Hybrid Embedded Feature Selection and Classification with Slim-TSF. arXiv:2409.04542.
[2] Posner, A. (2007). Up to 1-hour forecasting of radiation hazards from solar energetic ion events with relativistic electrons. Space Weather, 5(5), S05001.
[3] Torres, J. et al (2025). An ML approach to predicting SEP proton intensity and events using time series of relativistic electron measurements. Space Weather, 23(2), e2024SW003921.

This research is supported by the GSU Foundation Dean's Graduate Research Grant. We also gratefully acknowledge the Space Weather Workshop 2026 for providing the student support required to attend and present this work.