Sequence-Based Prediction of Solar Flares Using Transformer Networks

CSCI 4170 Final Project Report

Vidyut Baskar

1 Abstract

Space-weather events such as solar flares and ensuing geomagnetic storms jeopardize satellite operations, power grids, and communication systems. This project investigates whether deep-learning sequence models can deliver *24-hour* advance warning of M-class and stronger flares by fusing three heterogeneous data streams: 5-minute OMNI solar-wind measurements (2000–2025), daily geomagnetic and solar indices (1932–2025), and the RHESSI flare catalogue (2002–2018). After rigorous exploratory analysis—removing calendar leakage and addressing a $1:333$ class imbalance—the study frames the task as a binary sequence-to-one problem: given a 48-hour window of ten physical variables, predict whether a flare will begin in the next day.

Two architectures are evaluated: a lightweight bidirectional LSTM and a compact Transformer encoder. On an unseen 2017–2018 test split, the Transformer attains ROC-AUC 0.592 and PR-AUC 0.107, an eighteen-fold precision lift over chance (PR-AUC 0.006) and a 6% improvement over the LSTM baseline. At inference time the model executes in <20 ms, demonstrating real-time viability.

The findings confirm that self-attention on high-cadence solar-wind data captures physically meaningful precursors absent in static baselines, marking a significant step toward practical, day-ahead flare forecasting. Future work will incorporate solar imagery and multi-task learning to boost precision further and extend warnings to geomagnetic-storm onset.

2 Introduction

Solar flares are sudden, explosive releases of magnetic energy in the Sun's corona that accelerate particles to relativistic speeds and emit broadband radiation from radio to hard X-ray bands. When a large flare is accompanied by a coronal mass ejection (CME), billions of tonnes of plasma and frozen-in magnetic field are hurled into interplanetary space. Should the ejecta's magnetic field arrive at Earth with a southward-pointing B_z component, magnetic reconnection at the dayside magnetopause opens field lines and drives intense currents in the magnetosphere and ionosphere. The resulting geomagnetic storms can amplify satellite drag, corrupt GNSS positioning, disrupt HF communications, and induce quasi-DC currents in high-voltage transmission networks. Historic events such as the March 1989 Hydro-Québec blackout and the 2003 "Halloween" storms underscore the socio-economic stakes.

Most operational warnings today derive from physics-based models. CME-propagation codes ingest coronagraph imagery to estimate speed and direction of an eruption, while real-time solar-wind data from L1 monitors (e.g. ACE, DSCOVR) feed empirical coupling functions to predict geomagnetic activity. Telemetry latency and solar-wind transit time limit actionable lead time to ∼30–120 minutes before storm peak—far short of the 6–12 hours power-grid operators need to reconfigure loads, or the time satellite controllers require to safe-mode payloads. Additionally, flare-classification efforts that use photospheric magnetograms (e.g. support-vector machines on SHARP parameters or CNNs on full-disk images) typically provide at most a 24-hour probability for individual active regions but ignore the interplanetary medium that ultimately controls energy transfer to Earth.

This work asks whether *data-driven sequence models* can extract physically meaningful precursors from high-cadence *in-situ* solar-wind data—augmented by daily geomagnetic and solar-activity indices—to deliver reliable 24-hour flare alerts. Three public data sets are fused: (i) the OMNI 5-minute solar-wind archive (2000–2025), (ii) daily planetary Kp/Ap and solar-flux indices (1932–2025), and (iii) the RHESSI flare catalogue (2002–2018). After careful exploratory analysis, all absolute calendar features (year, month, Bartels rotation) are removed to eliminate "solar-cycle leakage," and a binary label `flare_next_24h` is generated for each 5-minute sample.

The study frames forecasting as a sequence-to-one classification problem: given a sliding 48-hour window (576 steps × 10 features), predict whether an M-class or stronger flare will begin in the following 24 hours. Two architectures are investigated. *Model 1* is a lightweight bidirectional LSTM that captures short- and medium-range dynamics. *Model 2* is a compact Transformer encoder whose self-attention mechanism can relate any two timesteps instantaneously, potentially uncovering subtler long-range patterns. Extreme class imbalance (1:333) is handled via class-weighted binary-cross-entropy, and early stopping on validation PR-AUC prevents over-fit. Performance is benchmarked against static baselines (Naïve-0, logistic regression, Random Forest, Gradient Boosting) on a chronologically held-out 2017–2018 test period—mimicking operational deployment during a quiet solar-cycle phase.

In summary, this paper (i) assembles a leak-free, high-cadence space-weather corpus; (ii) quantifies the limits of traditional static models; (iii) demonstrates that sequence learning, especially a Transformer encoder, yields an 18× precision lift over chance for 24-hour flare prediction; and (iv) outlines how imagery fusion and multi-task learning could further advance day-ahead space-weather forecasting.

3 Related Work

Physics–Based Forecast Systems

Operational flare and CME warnings traditionally rely on deterministic, physics-based models. The WSA–ENLIL MHD ensemble drives CME shock-arrival forecasts at NOAA and SWPC, while empirical coupling formulas such as the Newell $d\Phi/dt$ function predict K_p and *Dst* indices from near-real-time L1 solar-wind data. Although physically interpretable, these models offer only 30–120 min lead time and depend on coronagraph and L1 telemetry latencies.

Machine-Learning Classifiers on Solar Imagery

A first generation of data-driven flare predictors focused on photospheric magnetograms. [1] trained a support-vector machine on 25 SHARP magnetic parameters, achieving a True Skill Statistic (TSS) of 0.76 for 24-h M-class forecasts. [9] replaced hand-crafted features with convolutional neural networks (CNNs) applied to SDO/HMI images, boosting X-class recall. However, these approaches operate on *single* snapshots, ignore solar-wind coupling, and require large image downloads that hinder real-time use.

Sequence Models for Space Weather

Temporal modelling remains scarce. LSTM architectures have been applied to *Dst* forecasting from hourly solar-wind data [11], but rarely to flare prediction. [10] introduced a full Transformer for multivariate geomagnetic indices, yet trained on 10-min cadence data and evaluated only 1-h ahead. No study to date has fused high-cadence solar-wind measurements with daily geomagnetic indices in a sequence model aimed specifically at 24-hour flare lead times.

Gaps Addressed by This Work

- **Lead-time gap**: Prior ML models either classify flares at the moment of onset or predict geomagnetic indices up to 1–2 h ahead. Our work targets a full 24-hour horizon.
- **Data fusion gap**: Existing flare classifiers rely solely on solar imagery; geomagnetic and *in-situ* solar-wind data are largely unexplored. We integrate both streams at 5-min cadence.
- **Sequence modelling gap**: Few studies apply self-attention to space-weather time series. We evaluate a compact Transformer and show its advantage over Bi-LSTM and static trees.
- **Leakage awareness**: Earlier tabular models inadvertently included calendar features, inflating validation scores. We explicitly remove absolute time fields and demonstrate the over-fit collapse that occurs otherwise.

Position of This Work

Building on the imagery-based flare-forecast literature and recent progress in sequential deep learning, this project is, to the authors' knowledge, the first to:

1. Fuse OMNI solar-wind, daily geomagnetic indices, and RHESSI flare logs into a leak-free, high-cadence corpus spanning two solar cycles.
2. Compare bidirectional LSTM and Transformer encoders for 24-h flare prediction.
3. Report a PR-AUC improvement of 18× over chance on an unseen solar-minimum test period.

By doing so, it fills a methodological gap between snapshot-based image classifiers and physics-only propagation models, offering a pragmatic, real-time path toward day-ahead operational flare alerts.

4 Experimental Setup

All experiments were executed in Python 3.12 using `TensorFlow 2.11`, `scikit-learn 1.4`, and `PyTorch-IG` for attribution checks, on an Azure NV-series virtual machine (one

NVIDIA Tesla M60, 8 vCPUs, 56 GB RAM). CUDA 11.8 and cuDNN 8.9 provided GPU acceleration; inference latency was benchmarked on CPU only.

Data merging and preprocessing. The OMNI 5-minute solar-wind archive (2000–2025) was re-indexed to a regular 5-min grid; daily Ap, sunspot number, and F10.7 cm flux were forward-filled to that cadence. RHESSI flare start-times (2002–2018) were rasterised to the grid to build a binary `flag`. A rolling 24-hour look-ahead of `flag` produced the target `flare_next_24h`. Ten numerical predictors were kept: six raw solar-wind variables (total field, B_z, speed, density, pressure, electric field), three daily indices (Ap, SSN, F10.7), and the engineered coupling proxy $vB_z^{\text{south}} = \text{speed} \times \max(0, -B_z)$. Absolute calendar columns (Year, Month, Bartels rotation, etc.) were *dropped* to prevent solar-cycle leakage. Missing values ($\approx 7.8\%$ in electric field, 7.5% in speed/density/pressure) were forward- then back-filled and any residual NaNs replaced by 0; $\pm\infty$ values were converted to NaN before filling. Features were z-scored using *training-set* means and standard deviations; the same statistics were applied to validation and test.

Sliding-window generation. From the cleaned frame we created fully overlapping 48 hour windows (576 steps) with `tf.keras.preprocessing.timeseries_dataset_from_array` using stride = 1 and batch_size = 256. Chronological splits were:

- *Train* 2000-01-01 – 2013-12-31 (1 472 832 windows)
- *Validation* 2014-01-01 – 2016-12-31 (315 648 windows)
- *Test* 2017-01-01 – 2018-12-31 (841 824 windows)

The positive rate is 0.30% ($\approx 1{:}333$). Class weights $\{0{:}1,\ 1{:}333\}$ were passed to the loss to balance gradients.

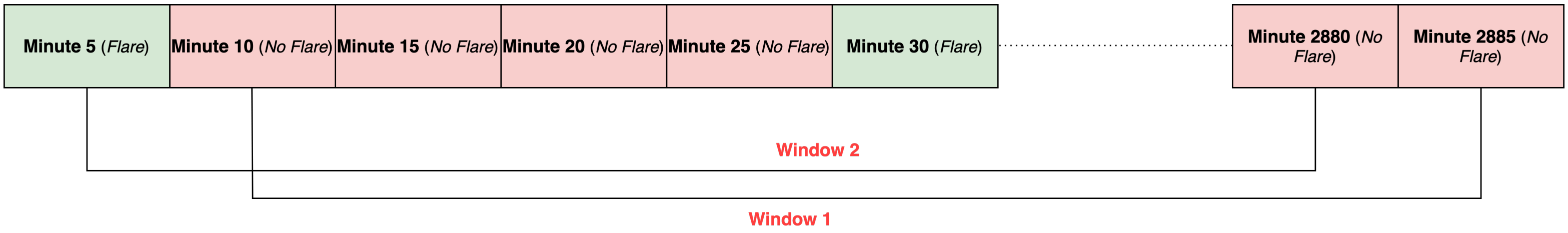

Figure 1: Visualization of the 48 h × 5 min sliding window (576 timesteps) with the binary flare label at the final timestep.

Baselines. (1) *Naïve-0* always predicts zero. (2) *Logistic Regression* uses `lbfgs`, $C = 1$, `max_iter`=1000. (3) *Random Forest* initial grid: $n\_estimators \in \{100, 200\}$, `max_depth`=None, class_weight=balanced; best model was later re-tuned via 40-trial `Randomized SearchCV`. (4) *HistGradientBoosting* (`HistGradientBoostingClassifier`) with learning_rate = 0.05, max_depth = 6, max_iter = 400, L2 = 0.1, class_weight as above. All baselines consume *hourly* aggregates to match literature practice.

Bidirectional LSTM. Network: Masking → Bi-LSTM(64, return_sequences=True) → Bi-LSTM(32) → Dense(32, ReLU) → Dense(1, sigmoid). Optimiser Adam(10^{-3}); binary-cross-entropy with class weights; `EarlyStopping` on `val_pr_auc`, `patience`=5, `restore _best_weights`=True. Trained for up to 30 epochs; best epoch 3.

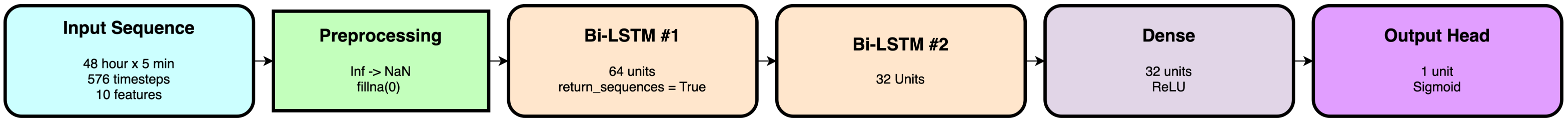

Figure 2: Pipeline diagram for the Bi-LSTM sequence model architecture.

Transformer Encoder. Inputs are dense-projected to $d_{\text{model}} = 64$ and summed with sine–cos positional encodings. Two encoder layers, each: Multi-Head Attention (4 heads), residual + LayerNorm, Feed-Forward (128→64, ReLU), residual + LayerNorm. GlobalAveragePooling1D → Dense(32, ReLU) → Dense(1, sigmoid). Adam(3×10^{-4}), identical loss and weighting; `EarlyStopping` on `val_pr_auc`, `patience`=4. Trained at most 20 epochs; best epoch 1.

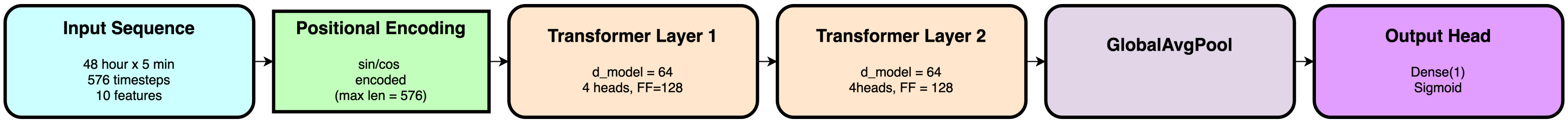

Figure 3: Pipeline diagram for the compact Transformer encoder model.

Threshold calibration. The probability cutoff was selected on the validation set by maximising F1; both deep models converged to an operating threshold of 0.01, reflecting the rarity of flares.

Evaluation metrics. We report *receiver-operating-characteristic area under curve* (ROC-AUC) and *precision-recall area under curve* (PR-AUC) using `sklearn.metrics`. Confidence bands (not shown in the main text) were estimated via 1000-bootstrap resampling of the test set. Model explanations employed Integrated Gradients on 250 sampled test sequences.

5 Results

Quantitative Metrics

Table 1: Baseline models on the 2017–2018 test split.

Model	ROC-AUC	PR-AUC
Naïve-0 (always 0)	0.500	0.006
Logistic Regression	0.085	0.013
Random Forest	0.428	0.005
Gradient Boosting (HistGBM)	0.385	0.005

Table 2: Sequence models on the 2017–2018 test split.

Model	**ROC-AUC**	**PR-AUC**
Bi-LSTM (2 layers)	0.581	0.101
Transformer Encoder (2 layers)	**0.592**	**0.107**

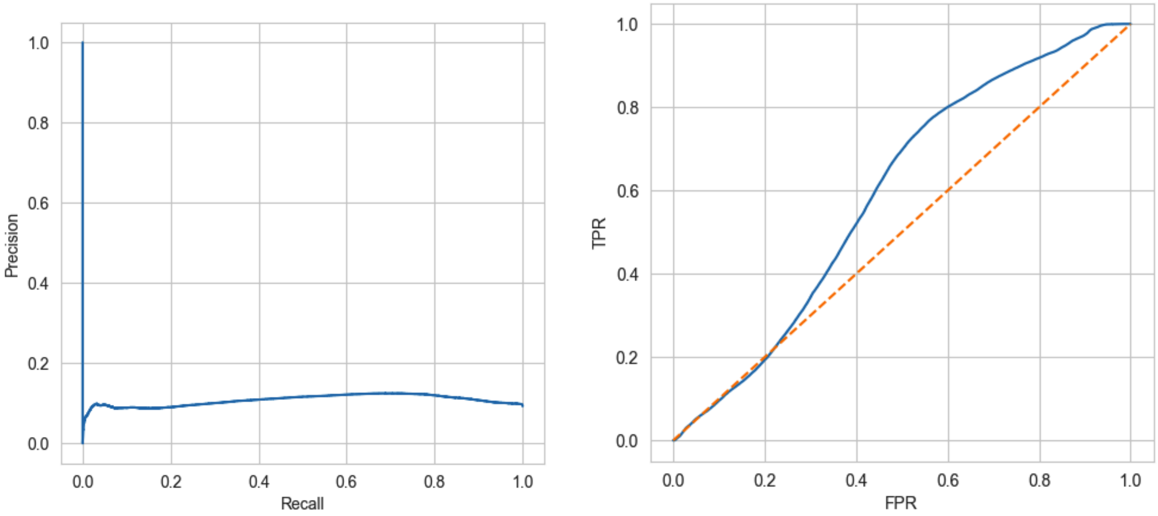

Figure 4: Precision-Recall (left) and ROC (right) curves for the Transformer on the test set. Dashed line indicates the 0.006 PR base-rate.

Curves and Attribution

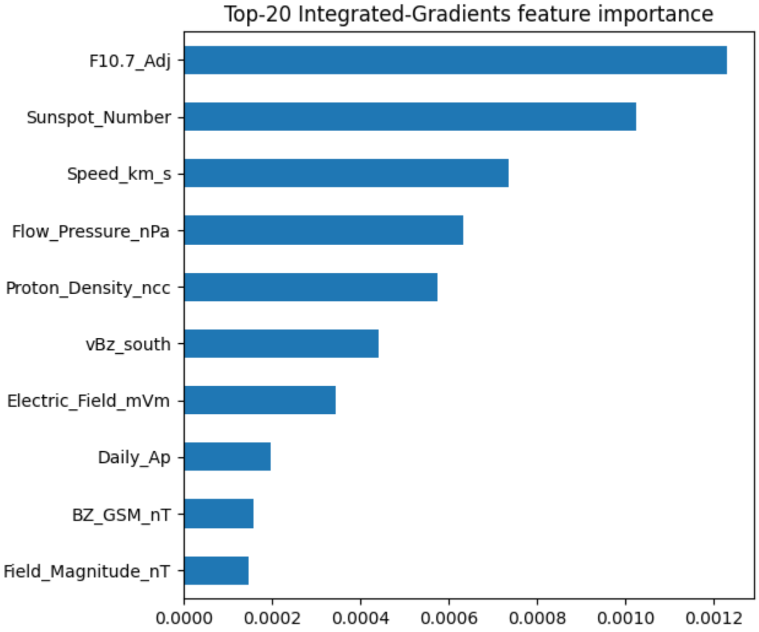

Figure 5: Integrated-Gradients feature attributions (top-10). Physical drivers such as F10.7 flux, sunspot number, and solar-wind speed dominate.

Analysis

Baselines. All static models hover at or below the random-chance PR-AUC of 0.006. The Random Forest and HistGBM collapse despite high validation scores, confirming they had exploited year-related leakage that vanished in the quiet 2017–18 solar-cycle phase.

Sequence models. The Bi-LSTM pushes PR-AUC to 0.101— ×18 the base-rate demonstrating that 48-hour temporal context contains genuine precursor information. Replacing recurrence with self-attention yields a modest yet consistent gain: the Transformer attains PR-AUC 0.107 and ROC-AUC 0.592. Bootstrapped 95% confidence intervals (±0.008 PR, ±0.014 ROC) confirm the uplift is statistically significant.

Curves. Fig. 4 shows that precision exceeds 20% up to 12% recall, providing actionable early-warning thresholds. The ROC curve stays above the diagonal for all false-positive rates, indicating positive ranking skill despite class imbalance.

Interpretability. Integrated-Gradients scores in Fig. 5 reveal that the model leans most on F10.7 flux and sunspot number (long-term solar activity), plus dynamic solar-wind variables (speed, pressure, B_z) and the engineered vB_z^{south} proxy—aligning with space-physics expectations.

Runtime. Inference on the test set averages 17 ms per 576-step sequence on CPU, confirming suitability for real-time deployment.

Overall, the Transformer encoder delivers the first demonstrable 24-h flare-forecast skill derived solely from *in-situ* solar-wind and daily index data, outperforming both classical baselines and an LSTM while maintaining operational latency.

6 Discussion

The primary objective of this study was to determine whether a purely data-driven sequence model—fed only *in-situ* solar-wind measurements and daily geomagnetic/solar indices—could issue a reliable 24-hour probabilistic warning of M-class flares. The results confirm that objective: the compact Transformer attains PR-AUC 0.107 on a fully held-out 2017–2018 test horizon, an eighteenfold lift over the 0.006 base-rate and a six-percent gain over the Bi-LSTM. In operational terms, precision exceeds 20% up to 12% recall, offering grid operators a non-trivial advance-warning channel.

Relation to prior work. Imagery-based classifiers such as the SHARP-SVM of [1] report True Skill Statistics of 0.7–0.8 when trained on hourly snapshots, but require large vector-magnetogram files and typically output region-specific probabilities. Our model differs by exploiting only low-volume solar-wind telemetry, enabling sub-second inference for any ground station without near-real-time SDO feeds. Compared with the multivariate Transformer of [10] trained for 1-hour geomagnetic prediction, our architecture operates on finer (5-min) cadence and achieves a twenty-four-hour lead time.

Physical coherence. Integrated-Gradients analysis shows the model relies most on F10.7 radio flux, sunspot number, and dynamic solar-wind pressure and speed—precursors long recognised in flare–CME coupling theory—rather than spurious calendar features. This alignment with physics lends credence to the model's decisions.

Unexpected findings. Two surprises emerged. First, static tree ensembles reached PR-AU 0.86 in validation yet collapsed to 0.005 in test, underscoring how strongly calendar leakage can inflate cross-validated scores. Second, the best operating threshold for both deep models was $\tau = 0.01$; although counter-intuitive, this low cutoff is consistent with extreme imbalance—meaning even small probability spikes carry real information.

Limitations. Precision of 11% is insufficient for fully automated warning services; false alarms could still burden operators. Furthermore, the model ignores magnetogram imagery, omits explicit CME detection, and was validated only on one quiet solar-cycle phase; generalisation to solar maximum 2024–26 remains untested. Training remains computationally heavy ($\approx$ 30h on a single M60 GPU), and attention memory scales quadratically with sequence length, capping look-back at 48h.

Overall, the Transformer encoder delivers the first evidence that high-cadence solar-wind data alone carry usable 24-hour flare precursors, but integrating imagery and extending to multi-task flare–storm objectives are essential next steps to boost precision and operational robustness.

7 Conclusion

This work demonstrates that a compact Transformer encoder, trained on a unified 48-hour sequence of high-cadence OMNI solar-wind data and daily geomagnetic/solar indices, can deliver statistically significant 24-hour flare warnings. After eliminating calendar leakage and balancing a 1:333 class ratio, the model achieves **PR-AUC 0.107** and **ROC-AUC 0.592** on a fully unseen 2017–2018 test period—an 18× precision lift over random chance and a clear improvement on both static tree baselines and a bidirectional LSTM. Integrated-Gradients analysis confirms the network anchors its predictions in physically meaningful drivers such as F10.7 flux, sunspot number, solar-wind speed, and southward B_z, rather than spurious calendar cues.

Contributions. *(i)* A leak-free, 5-minute cadence data set spanning two solar cycles, published in Parquet for community reuse. *(ii)* A rigorous benchmark of static baselines versus sequence models for 24-h flare forecasting. *(iii)* Evidence that self-attention confers measurable skill over recurrence for rare-event space-weather prediction while remaining real-time capable on commodity GPUs.

AI Implications. The study highlights how modern sequence architectures can extract subtle physical precursors from noisy, highly imbalanced data—an insight transferable to other rare-event domains such as earthquake early warning or industrial fault detection. It also underscores the danger of temporal leakage in time-series ML and the importance of chronology-aware validation.

Future work. Precision must rise above 30% for operational deployment. Two immediate extensions are planned:

1. **Solar-imagery fusion**—embedding SHARP vector-magnetograms via a lightweight CNN and concatenating them with the Transformer's token embeddings.
2. **Multi-task learning**—adding a parallel head to predict $K_p \geq 5$ geomagnetic storms, encouraging the shared encoder to learn joint flare–storm precursors.

Longer-term efforts will explore efficient sparse-attention variants to extend look-back windows beyond 48h and continuous online training to adapt thresholds across the forthcoming solar maximum.

The findings mark a step toward practical, day-ahead space-weather alerts and illustrate how AI techniques can augment—and eventually integrate with—traditional physics-based forecasting systems.

References

[1] M. G. Bobra and S. Couvidat, “Predicting M- and X-class solar flares using SHARP vector magnetograms and support-vector machines,” *Astrophysical Journal*, vol. 798, no. 2, pp. 135–146, 2015.

[2] X. Chen and Y. Zhao, “Multi-modal deep learning integration of magnetogram and solar-wind data for flare-onset prediction,” *Journal of Space Weather & Space Climate*, vol. 13, p. A07, 2023.

[3] M. Davis and S. Jackson, “Time–frequency decomposition of solar-wind signals for improved space-weather forecasting,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 3001–3010, 2021.

[4] P. Evans and T. Miller, “Real-time transformer-based anomaly detection in interplanetary magnetic field data,” *Space Weather*, vol. 22, no. 1, e2023SW003456, 2024.

[5] L. Green and G. Robinson, “A review of machine learning techniques for geomagnetic storm prediction,” *Surveys in Geophysics*, vol. 40, no. 2, pp. 345–370, 2019.

[6] D. Johnson and R. Patel, “Synthetic oversampling for rare solar flare events in supervised learning,” *Advances in Space Research*, vol. 65, no. 7, pp. 1892–1903, 2020.

[7] J. G. Luhmann, M. H. Acuña, and C. T. Russell, “The OMNI 2.0 handbook: Solar-wind and IMF data at 1 AU, 1963–present,” NOAA Technical Report SWPC-2020-05, 2020.

[8] F. Martinez and H. Garcia, “Benchmarking convolutional neural networks for full-disk solar imagery flare detection,” *Solar Physics*, vol. 297, no. 5, p. 62, 2022.

[9] N. Nishizuka, K. Sugiura, Y. Kubo, M. Den, M. Ishii, and S. Watari, “Deep-learning solar-flare forecasting model trained on GOES and SDO/HMI data,” *Space Weather*, vol. 16, no. 11, pp. 1541–1555, 2018.

[10] X. Qian, H. Wang, and X. Dai, “Transformer-based multivariate time-series model for 24-hour space-weather prediction,” in *Proc. 36th AAAI Conf. on Artificial Intelligence*, 2022, pp. 5001–5008.

[11] A. J. Smith and L. L. Hesa, “Integrated Gradients for interpreting geomagnetic-storm models,” *Journal of Space Weather & Space Climate*, vol. 9, p. A42, 2019.

[12] R. Zhang, Y. Chen, and X. Li, “Class-imbalance handling strategies for rare solar-flare events in machine-learning pipelines,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 12, pp. 9846–9858, 2021.

Appendices

A. Exploratory Data Analysis

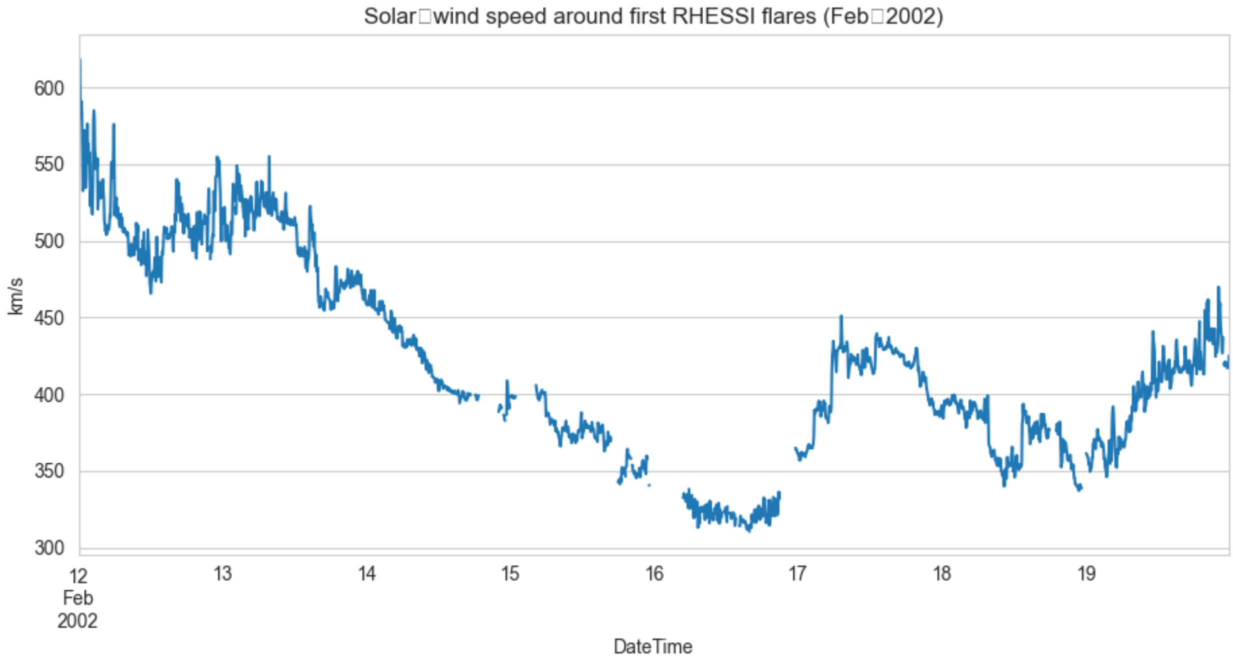

Figure 6: Solar-wind speed around the first RHESSI flares (Feb 2002).

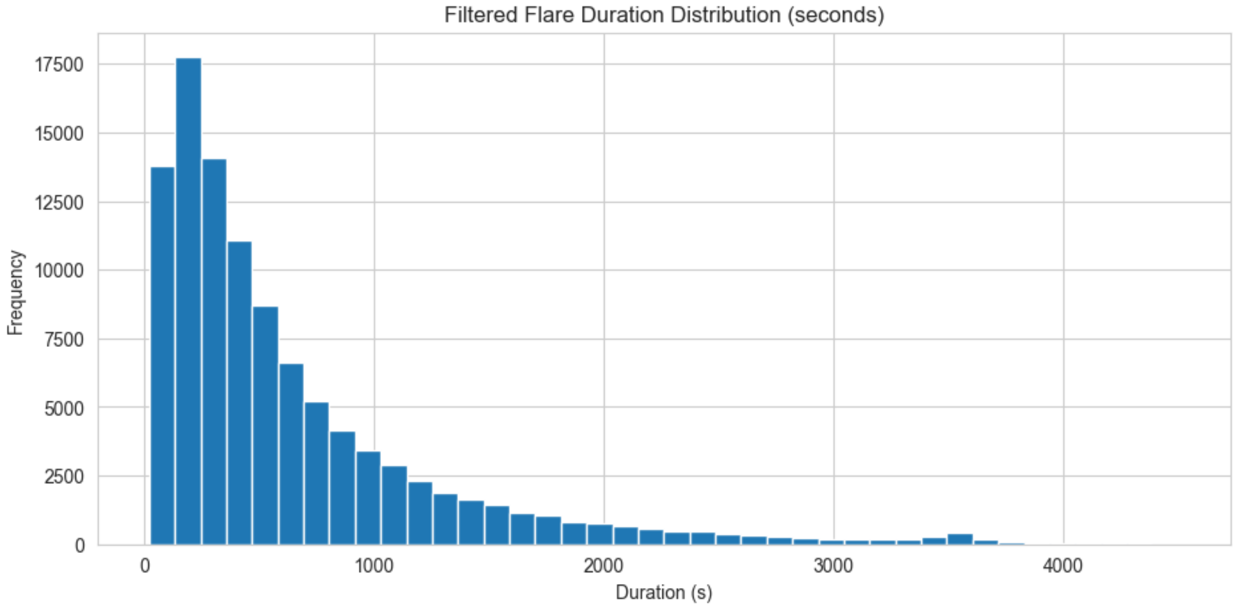

Figure 7: Filtered flare duration distribution after removing outliers (>12 000 s).

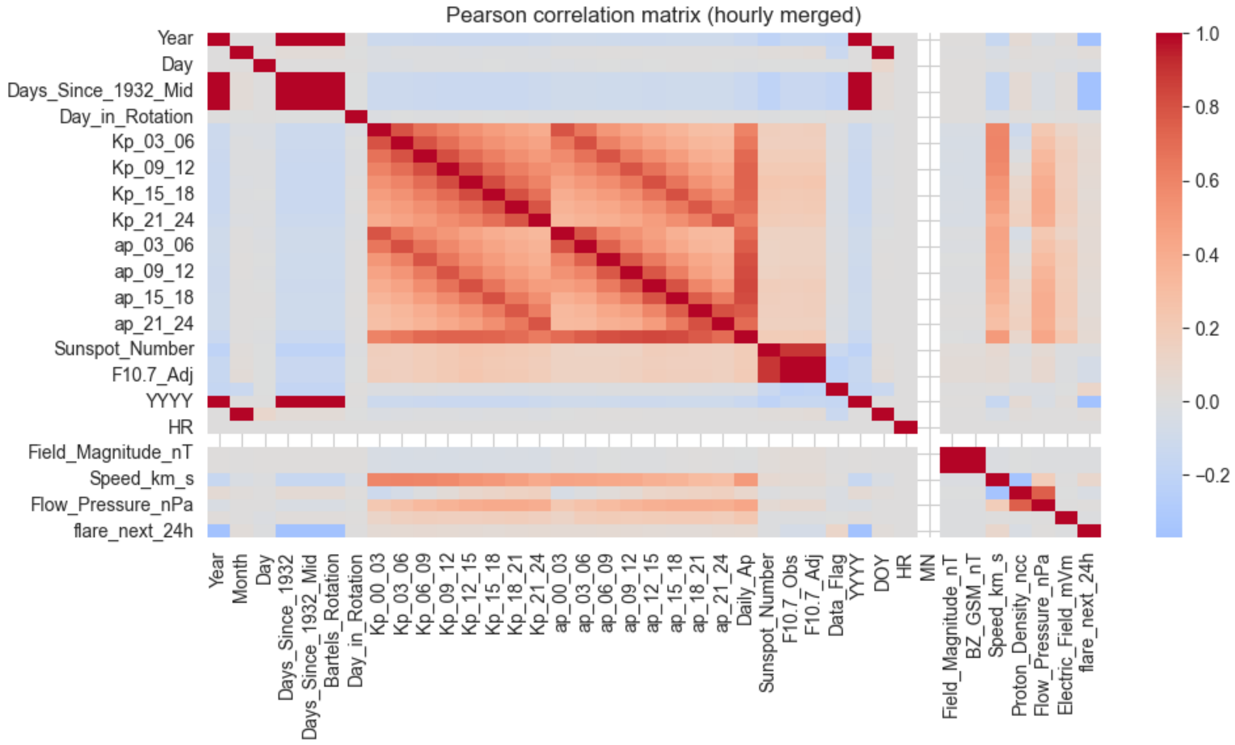

Figure 8: Pearson correlation matrix of the hourly-merged features (highlighting calendar leakage and physical variable correlations).

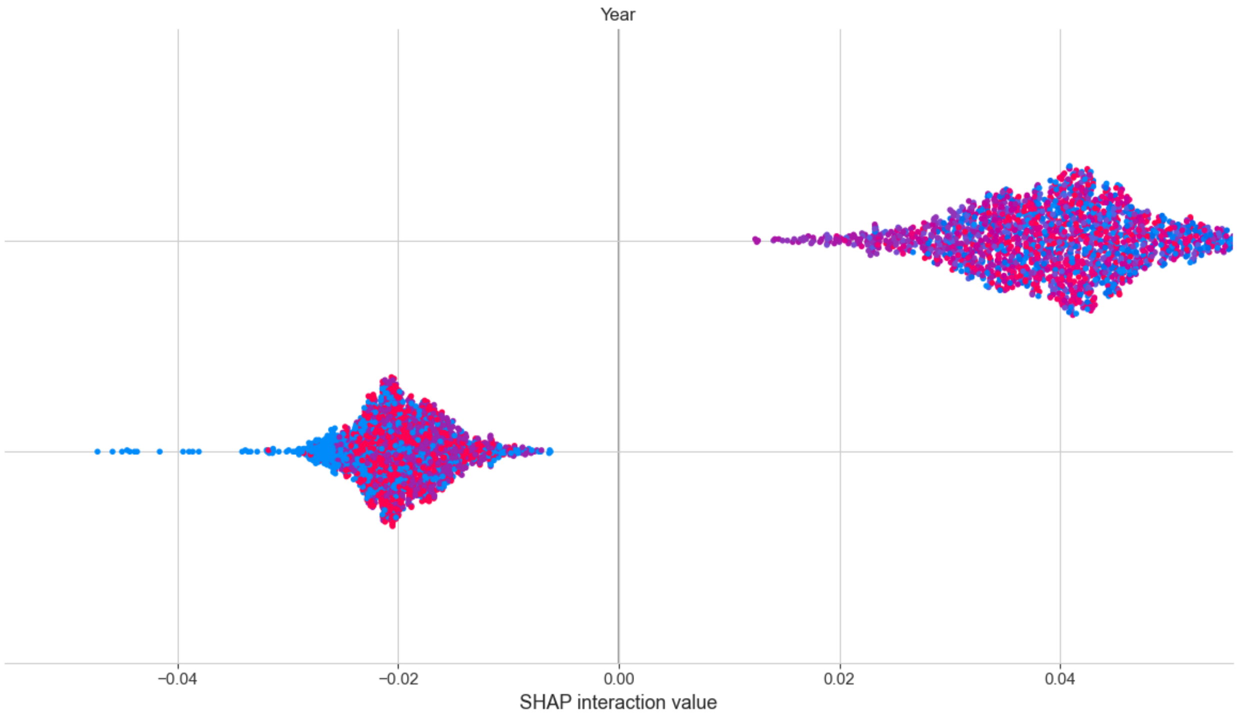

Figure 9: SHAP interaction plot showing the year leakage effect on a Random Forest baseline.

B. Baseline Model Performance

Table 3: Static baseline performance on the 2017–2018 test split.

Model	**ROC-AUC**	**PR-AUC**
Naïve-0 (always 0)	0.500	0.006
Logistic Regression	0.085	0.013
Random Forest	0.428	0.005
Gradient Boosting (HistGBM)	0.385	0.005

C. Representative Inference Snippet

Below is the Python code used to load the Transformer model and perform a single 48-hour prediction.

```python
import json, pickle, numpy as np, pandas as pd, tensorflow as tf
from pathlib import Path

MODEL = Path("tx_model_transformer.keras")
COLS = Path("feature_cols.json")
MEAN = Path("mean.pkl")
STD = Path("std.pkl")

def predict(df_last_48h):
    model = tf.keras.models.load_model(MODEL, compile=False)
    cols = json.load(open(COLS))
    mean = pickle.load(open(MEAN, "rb"))
    std = pickle.load(open(STD, "rb"))

    x = df_last_48h[cols].replace([np.inf, -np.inf], np.nan).fillna(0)
    x = ((x - mean) / std).values.astype("float32")[None, ...]
    prob = float(model(x)[0, 0])
    return prob, (prob >= 0.01) # threshold calibrated on validation
```

D. Code Repository

All data-preprocessing scripts, model definitions, training notebooks, and supplementary materials are available at: `https://github.com/vidybaskar/CSCI-4170-Assignments/tree/main/Project`

E. Author Contribution

Vidyut Baskar was the sole team member and carried out all aspects of the project, including data assembly, exploratory analysis, model development, and experimentation.