Using Deep Learning Techniques to Predict Solar Switchbacks

Joaquín Robles Riaza

September 14, 2025

Abstract

Switchbacks are sudden, large deflections of the interplanetary magnetic field, and represent one of the most important discoveries of NASA's Parker Solar Probe (PSP). Their origin and potential role in solar wind heating remain an open question in heliophysics. This project aims to study whether short sequences of PSP data can be used to predict the occurrence of switchbacks using recurrent neural network architectures such as LSTMs or GRUs. The first phase focuses on obtaining data from the FIELDS and Solar Probe Cup (SPC) instruments on PSP to reconstruct a switchback catalogue. The second phase analyzes the data and builds LSTM and GRU architectures to predict switchbacks from short sliding windows. Preprocessing strategies, class imbalance, and threshold definitions were systematically addressed. Despite the difficulty of the task, the methodology establishes a reproducible baseline and provides valuable insight into the challenges of forecasting switchbacks and space weather.

1 Introduction

The Parker Solar Probe (PSP) was launched in 2018 and has since provided unprecedented measurements of the solar wind close to the Sun. One of its major discoveries is *switchbacks*, which are sudden and large rotations of the solar wind's magnetic field vector. These rare and extreme events were first reported during PSP's first perihelion encounter.

It is thought that switchbacks play a crucial role in space weather, particularly in solar wind acceleration and heating, as they are associated with bursts of Alfvénic turbulence and intermittent reconnection-like structures [1]. Switchbacks are of great scientific interest mainly because:

- They represent a key dynamical process in the young solar wind.
- Their origin is still debated (generation in the corona or just solar wind instabilities).
- Predicting their occurrence could provide a window into turbulent energy transfer mechanisms.

This project aims to address whether it is possible to predict the onset of switchbacks from very short PSP time sequences using LSTMs and GRUs.

2 Data Acquisition and Switchback Catalogue Construction

The entire data processing pipeline described below in Section 2, together with the corresponding implementation code, is available in a Jupyter Notebook which can be accessed at the following link:
`psp_data_and_switchback_catalogue.ipynb`

2.1 Data Sources

The primary dataset employed in this study was obtained from NASA's Parker Solar Probe (PSP) mission. The data were accessed predominantly through the `pyspedas` library, which provides standardized tools to retrieve and preprocess heliophysics mission data. Two PSP instruments were used in this work:

- **Level 2 FIELDS:** Vector measurements of the interplanetary magnetic field. The dataset includes the three RTN components (B_R, B_T, B_N), which describe the radial, tangential, and normal magnetic field components relative to the spacecraft's position with respect to the Sun.
- **Level 3 SPC:** Solar wind plasma parameters derived from the Solar Probe Cup (SPC). These parameters include proton and alpha particle velocity distribution fits, providing estimates of density, thermal speed, and the three components of the bulk velocity vector.

In addition to these datasets, the spacecraft's heliocentric distance was obtained via the `sunpy` library. This distance information, expressed in astronomical units (AU), provides contextual information regarding PSP's orbit and radial position during the selected study interval.

The time interval chosen spans November 1 to November 10, 2018, covering the first perihelion passage of PSP on November 6. This interval was selected for two primary reasons:

- Switchbacks occur with higher frequency close to the Sun, typically within 0.2 AU, making this perihelion encounter particularly relevant for their study.
- Both FIELDS and SPC instruments operated with increased cadence during this interval, enabling a high-resolution reconstruction of the solar wind plasma state and magnetic field deflections.

2.2 Construction of the Switchback Catalogue

The raw Level 2 FIELDS dataset, containing the three vector components of the magnetic field, exhibited very fine temporal resolution. Its irregular sampling cadence ranged between 0.00334 s and 0.01387 s, yielding approximately 151 million measurements over the selected 10-day interval. Although this granularity is valuable for wave and turbulence studies, it represents an impractically large volume of data for the purposes of switchback catalog construction and predictive modeling. To balance computational tractability with physical fidelity, the magnetic field measurements were downsampled to a constant 5-second cadence by averaging all measurements within each non-overlapping 5-second interval. This resampling procedure effectively suppressed high-frequency noise, reduced the data size to 155,520 uniformly spaced rows, and retained the mesoscale magnetic field variations relevant for switchback identification.

The Level 3 SPC dataset, consisting of proton and alpha particle velocity distribution fits, was also provided at an irregular high-frequency cadence ranging from 0.0512 s to 0.8738 s. Similar to the FIELDS data, it was resampled to a uniform 5-second cadence.

The following proton fit parameters were preserved for analysis:

- $n_{p,\mathbf{fit}}$**:** Proton number density, quantifying the concentration of protons in the solar wind, expressed in particles per cm^3.

- $w_{p,\mathbf{fit}}$**:** Proton thermal speed, corresponding to the thermal broadening of the proton velocity distribution and related to proton temperature.

- $v_{p,\mathbf{fit},R}$**:** Radial component of the proton bulk velocity, V_R, directed outward along the Sun–spacecraft line.

- $v_{p,\mathbf{fit},T}$**:** Tangential component of the proton bulk velocity, V_T, aligned with the solar equatorial plane and azimuthal direction.

- $v_{p,\mathbf{fit},N}$**:** Normal component of the proton bulk velocity, V_N, oriented perpendicular to both R and T and completing the RTN coordinate system.

The spacecraft distance data had an original granularity of one measurement every 2 minutes. To maintain temporal consistency with the resampled FIELDS and SPC datasets, the distance time series was also resampled to a 5-second cadence. This procedure introduced NaN values at intermediate time steps between the original 2-minute measurements, which were later handled.

Once resampled, the FIELDS, SPC, and distance datasets were merged into a single combined dataframe. The alignment was performed using the FIELDS timestamps, which were complete and contained no missing entries. This ensured that all plasma and positional parameters were consistently aligned with magnetic field measurements at a common 5-second cadence. This process provided a robust and homogeneous dataset for switchback detection.

Switchbacks were identified using a modified version of the methodology introduced by [2]. Unlike the original fixed-window method, which computed magnetic field deflections relative to a mean field in discrete 3- or 6-hour blocks, a moving-window approach was employed. For each time step, the instantaneous magnetic field direction was compared against the local mean field computed over a symmetric 3-hour rolling window (i.e., 1.5 hours before and 1.5 hours after the central point). This approach provided continuous detection throughout the dataset and eliminated artifacts associated with discrete block boundaries.

The procedure can be summarized as follows:

1. **Local background magnetic field:** A centered rolling average of 3 hours (2160 samples at 5-second cadence) was applied to each of the

three components B_r, B_t, B_n to obtain the background magnetic field:

$$\mathbf{B}_{\text{avg}}(t) = \big(B_{R,\text{avg}}(t), B_{T,\text{avg}}(t), B_{N,\text{avg}}(t)\big).$$

2. **Deflection angle:** Both the instantaneous magnetic field vector and its corresponding local background vector were normalized to unit length:

$$\hat{\mathbf{B}}(t) = \frac{\mathbf{B}(t)}{\|\mathbf{B}(t)\|}, \quad \hat{\mathbf{B}}_{\text{avg}}(t) = \frac{\mathbf{B}_{\text{avg}}(t)}{\|\mathbf{B}_{\text{avg}}(t)\|}.$$

The cosine of the deflection angle $\alpha(t)$ between these vectors was then computed as:

$$\cos\alpha(t) = \hat{\mathbf{B}}(t) \cdot \hat{\mathbf{B}}_{\text{avg}}(t).$$

From this, the deflection angle itself follows:

$$\alpha(t) = \arccos\Big(\hat{\mathbf{B}}(t) \cdot \hat{\mathbf{B}}_{\text{avg}}(t)\Big),$$

constrained to the interval $\alpha \in [0°, 180°]$.

3. **Switchback criterion:** A time step was classified as a switchback if the magnetic field deflection angle satisfied the condition:

$$\alpha(t) > 90°.$$

3 Data Analysis and Modeling

To ensure full reproducibility of the results, the complete workflow described in Section 3 (*Data Analysis and Modeling*), including both *Data Preparation and Preprocessing* and *Predictive Modeling*, has been implemented in a Jupyter Notebook. This notebook contains all relevant code, intermediate outputs, and detailed steps necessary to replicate the analyses presented in this section, and can be accessed at the following link:
`psp_data_processing_and_modeling_for_switchback_prediction.ipynb`

3.1 Data Preparation and Preprocessing

3.1.1 Statistical Analysis and Missing Values

The dataset constructed in the previous phase was imported and systematically explored. The Level 3 SPC data exhibited missing values. Specifically, between 1.485% and 1.530% of the proton fit parameters contained gaps, while 100% of the alpha particle fits were missing (all entries were NaN).

Furthermore, the distance to the Sun also contained 95.83% NaN values given the granularity it originally had. The descriptive statistics of the raw dataset are presented in Table 1.

Variable	**Mean**	**Std**	**Min**	**Max**	**% NaN**
B_R	-52.01	26.63	-103.67	83.01	0.00
B_T	12.15	30.24	-92.92	108.21	0.00
B_N	-0.88	26.58	-87.16	93.86	0.00
$n_{p,\mathrm{fit}}$	283.68	121.58	45.19	9839.60	1.53
$w_{p,\mathrm{fit}}$	77.03	22.06	7.96	284.59	1.49
$v_{p,\mathrm{fit},R}$	335.62	73.83	19.77	729.59	1.49
$v_{p,\mathrm{fit},T}$	20.00	35.90	-169.92	122.80	1.53
$v_{p,\mathrm{fit},N}$	0.21	40.39	-219.85	346.69	1.53
$n_{a,\mathrm{fit}}$	NaN	NaN	NaN	NaN	100.00
$w_{a,\mathrm{fit}}$	NaN	NaN	NaN	NaN	100.00
$v_{a,\mathrm{fit},R}$	NaN	NaN	NaN	NaN	100.00
$v_{a,\mathrm{fit},T}$	NaN	NaN	NaN	NaN	100.00
$v_{a,\mathrm{fit},N}$	NaN	NaN	NaN	NaN	100.00
$\mathrm{distance}_{\mathrm{au}}$	0.19	0.02	0.17	0.24	95.83
switchback	0.03	0.16	0.00	1.00	0.00

Table 1: Statistical summary of the variables, including mean, standard deviation, minimum, maximum, and percentage of missing values.

Given these limitations, all alpha particle fit variables were discarded from further analysis. For the proton fits, rather than discarding rows containing missing values—which would have introduced bias and reduced temporal coverage—the gaps were imputed using a k-Nearest Neighbors (KNN) imputer. This method leverages similarities between neighboring samples in the high-dimensional feature space to infer plausible replacements, thereby preserving both temporal continuity and physical coherence.

For the variable measuring distance to the Sun, $distance_{au}$, since the distance must be a continuous function, the NaN values were filled using linear interpolation between the known data points.

After these processes, no missing values remained in the dataset. The descriptive statistics of the cleaned dataset, after alpha particle fit variable removal, KNN imputation of proton fits, and linear interpolation of missing PSP distance values, are shown in Table 2.

Variable	Mean	Std	Min	Max	% NaN
B_R	-52.01	26.63	-103.67	83.01	0.0
B_T	12.15	30.24	-92.92	108.21	0.0
B_N	-0.88	26.58	-87.16	93.86	0.0
$n_{p,\mathrm{fit}}$	284.11	120.96	45.19	9839.60	0.0
$w_{p,\mathrm{fit}}$	77.02	21.96	7.96	284.59	0.0
$v_{p,\mathrm{fit},R}$	335.55	73.54	19.77	729.59	0.0
$v_{p,\mathrm{fit},T}$	20.05	35.85	-169.92	122.80	0.0
$v_{p,\mathrm{fit},N}$	0.02	40.37	-219.85	346.69	0.0
Distance (au)	0.19	0.02	0.17	0.24	0.0
Switchback	0.03	0.16	0.00	1.00	0.0

Table 2: Descriptive statistics (mean, standard deviation, min, max, and percentage of missing values) for each variable after preprocessing.

3.1.2 Dataset Partitioning

To enable model training and evaluation, the dataset was divided into training, validation, and test subsets. Since the data are time series, the partitioning was performed along the temporal axis to prevent information leakage from the future into the past. The split was defined as follows:

- **Training set:** First 60% of the dataset.
- **Validation set:** Next 20% of the dataset (from 60% to 80%).
- **Test set:** Final 20% of the dataset.

3.1.3 Distribution Analysis and Scaling

Each magnetic field component and solar wind proton fit variable from the training set was examined to characterize its distribution. Histograms with fitted curves were generated (Figure 1), and skewness and kurtosis were calculated for each parameter.

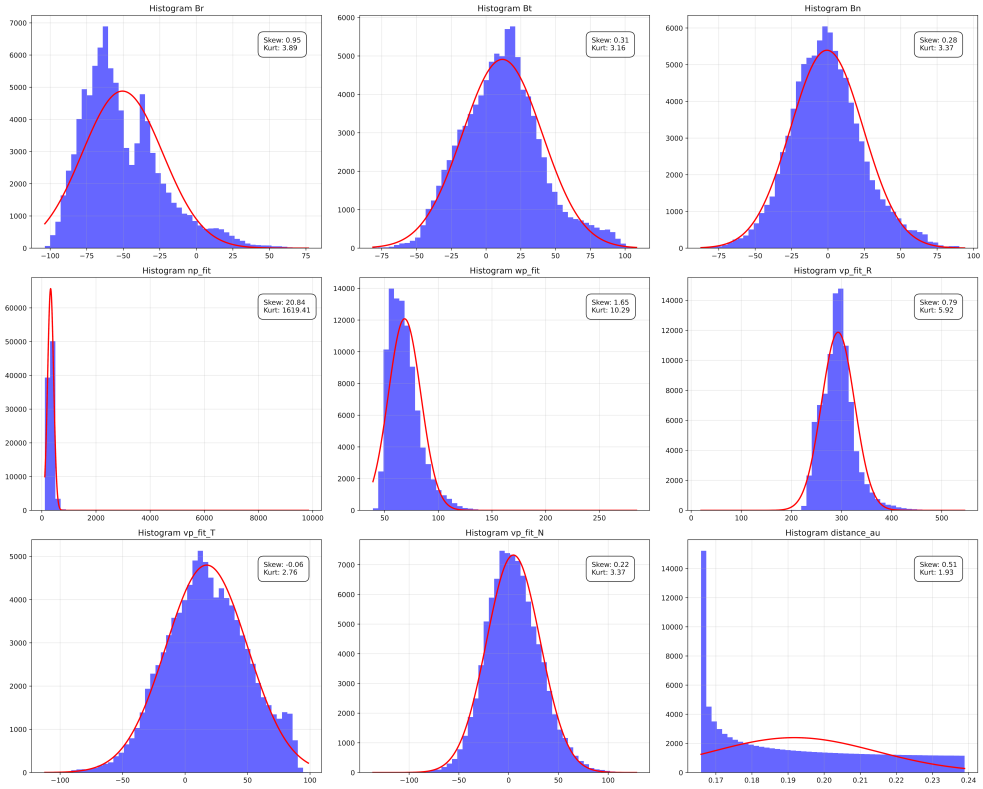

Figure 1: Histograms of magnetic field components, solar wind proton fit variables, and distance to the Sun, with corresponding skewness and kurtosis values.

Following established criteria, a variable was considered approximately Gaussian distributed if:

- The skewness lay between -0.5 and 0.5, indicating approximate symmetry.
- The kurtosis lay between 2.5 and 3.5, indicating a shape close to the normal distribution.

Table 3 presents the computed skewness, kurtosis, and Gaussian classification for each variable.

Variable	Skewness	Kurtosis	Classification
B_R	0.95	3.90	Non-Gaussian
B_T	0.31	3.16	Gaussian
B_N	0.28	3.37	Gaussian
$n_{p,\mathrm{fit}}$	20.84	1619.41	Non-Gaussian
$w_{p,\mathrm{fit}}$	1.65	10.29	Non-Gaussian
$v_{p,\mathrm{fit},R}$	0.79	5.92	Non-Gaussian
$v_{p,\mathrm{fit},T}$	-0.06	2.76	Gaussian
$v_{p,\mathrm{fit},N}$	0.22	3.37	Gaussian

Table 3: Skewness, kurtosis, and Gaussian classification of variables.

Based on these thresholds, the variables identified as approximately Gaussian were B_T, B_N, $v_{p,\mathrm{fit},T}$, and $v_{p,\mathrm{fit},N}$. The remaining variables—B_R, $n_{p,\mathrm{fit}}$, $w_{p,\mathrm{fit}}$, and $v_{p,\mathrm{fit},R}$—were classified as non-Gaussian.

Accordingly, preprocessing transformations were selected as follows:

- Gaussian-distributed variables were standardized using a **StandardScaler**, which centers the data and scales it to unit variance.
- Non-Gaussian variables were normalized using a **MinMaxScaler**, which maps the values to a fixed range while preserving their distributional shape.

This dual-scaling strategy ensures that each feature is transformed in a manner consistent with its underlying distribution, facilitating robust model training [3].

3.2 Predictive Modeling

The final stage of the methodology focused on developing recurrent neural network (RNN) models capable of forecasting switchbacks from short sequences of solar wind and magnetic field measurements. This section details the creation of the input samples via sliding windows, the handling of class imbalance, the recurrent architectures tested, the comparative evaluation of model stability, threshold optimization and testing of the model on test data.

3.2.1 Windowing for Short Sequences

As outlined in Section 1, the primary objective of this study was to investigate whether the onset of switchbacks could be predicted from very short sequences of Parker Solar Probe data. To this end, the training dataset was segmented into overlapping windows of fixed length. Each input sample consisted of a 1-minute window of consecutive data points. Given the resampled cadence of 5 seconds, this corresponds to 12 time steps per sample.

Each window was labeled according to whether a switchback occurred in the subsequent time step. This formulation transforms the task into a binary classification problem in which the model must infer precursors to switchbacks rather than detect ongoing magnetic deflections, thereby significantly increasing the difficulty of the prediction task.

3.2.2 Loss Function and Class Weights

An analysis of the target variable revealed a severe imbalance in the dataset. Only approximately 2.71% of the time steps were labeled as switchbacks ($y = 1$), while the remaining 97.29% corresponded to non-switchback intervals ($y = 0$). Training without addressing this imbalance would strongly bias the model towards always predicting the majority class.

To mitigate this, class weights were incorporated into the binary cross-entropy loss function:

$$\mathcal{L} = -w_1\, y\, \log(\hat{y}) - w_0\, (1 - y)\, \log(1 - \hat{y}), \tag{1}$$

where y denotes the true label, $\hat{y}$ the predicted probability, and w_0, w_1 the weights associated with classes 0 and 1, respectively.

The weights were computed based on inverse class frequencies [4]:

$$w_c = \frac{2N}{N_c}, \tag{2}$$

where N is the total number of samples and N_c the number of samples belonging to class c. This scheme penalizes misclassifications of rare switchbacks more heavily, effectively rebalancing the training process.

For the dataset under study, the resulting class weights were:

$$w_0 \approx 0.51,$$
$$w_1 \approx 18.43.$$

This weighting strategy tried to ensure that the learning process gave appropriate importance to the minority class, enabling the models to capture predictive features associated with switchback occurrences.

3.2.3 Model Architectures and Comparative Evaluation

To capture the temporal signatures preceding switchback events, two families of recurrent neural network (RNN) architectures were explored and benchmarked in this study. The focus was on architectures specifically designed to model sequential dependencies, as the task required extracting predictive patterns from short consecutive windows of solar wind data.

- **Long Short-Term Memory (LSTM)** networks: LSTMs introduce input, forget, and output gating mechanisms that regulate the flow of information across time steps. This design mitigates the vanishing gradient problem of traditional RNNs, allowing the model to learn both short- and long-term dependencies. Given the potentially subtle and extended temporal signatures of switchbacks, LSTMs were considered a strong candidate for this task.
- **Gated Recurrent Unit (GRU)** networks: GRUs employ a simplified gating structure (reset and update gates) compared to LSTMs, thereby reducing the number of parameters while maintaining the ability to capture relevant temporal patterns. Their computational efficiency and reduced complexity make GRUs particularly suitable for scenarios where model stability and training time are critical.

For both architectures, a systematic hyperparameter exploration was performed, varying the following factors:

- **Network depth:** The number of recurrent layers was varied between 2 and 3. Increasing depth allows the network to model increasingly abstract temporal representations, though at the cost of greater computational demand and risk of overfitting.
- **Hidden dimension size:** Each recurrent layer was configured with a different number of neurons (ranging from 16 to 256). Larger hidden dimensions increase representational capacity but may also require stronger regularization to ensure generalization.
- **Dropout rate:** Dropout values between 0.1 and 0.2 were tested. By randomly deactivating a fraction of neurons during training, dropout

serves as an effective mechanism to reduce overfitting and improve robustness.

- **Activation function:** Hyperbolic tangent (`tanh`) activations was always used because it is traditionally employed in recurrent architectures due to its bounded output range [5].

All tested models were designed following a **pyramidal architecture**, where the number of hidden neurons decreases with depth (e.g., 128–64–32 or 64–32–16). This design choice is motivated by the need to progressively compress information into lower-dimensional, higher-level representations, reducing the risk of overfitting while encouraging feature abstraction. Pyramidal RNNs are widely adopted in sequence modeling tasks such as speech recognition and time-series forecasting due to their favorable trade-off between expressiveness and efficiency [6].

Given the strong imbalance in the dataset (see Section 3.4.2), all models were trained with class-weighted binary cross-entropy as the loss function. This ensured that rare switchback events contributed proportionally more to the optimization process, counteracting the bias towards the majority class.

To assess the stability and robustness of these models, a k-fold cross-validation strategy was employed on the validation set. In each fold, the model was trained from scratch to ensure independence of results. The **F_1-score** was adopted as the primary evaluation metric, as accuracy would be highly misleading under the extreme class imbalance present in this task. The F_1-score balances the trade-off between **precision** (the fraction of predicted switchbacks that are true switchbacks) and **recall** (the fraction of true switchbacks that were successfully detected). Formally, it is defined as the harmonic mean of precision and recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

This formulation ensures that both false positives and false negatives are penalized, making it a rigorous and informative measure of predictive quality in imbalanced binary classification problems such as switchback detection.

Among the large pool of tested architectures, the five best-performing models, identified based on their mean cross-validated F_1-scores, are reported in Table 4. Each model is described in terms of architecture type, depth, hidden dimension configuration, activation function, and dropout regularization.

Model	N. of Layers	Neurons per Layer	Activation	Dropout
Model 1	3	[256, 128, 64]	Tanh	0.1
Model 2	3	[64, 32, 16]	Tanh	0.1
Model 3	2	[128, 64]	Tanh	0.1
Model 4	2	[128, 64]	Tanh	0.2
Model 5	2	[256, 128]	Tanh	0.2

Table 4: Summary of the five best-performing recurrent neural network models, showing the number of layers, neurons per layer, activation function, and dropout rate.

The variability of performance across folds for these five models is illustrated in Figure 2, which shows the distribution of F_1-scores via boxplots. This representation highlights both the median predictive power and the stability of each model under different validation splits.

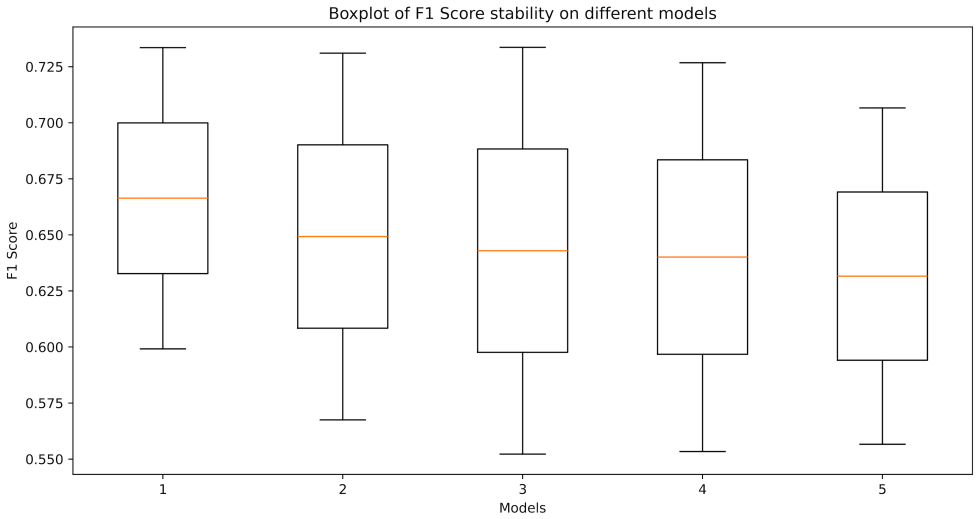

Figure 2: Distribution of cross-validated F_1-scores for the five best-performing models. Boxplots show variability across folds.

The comparative evaluation revealed that **Model 1** not only demonstrated the most favorable average F_1-score but also the highest robustness. This model, consisting of three stacked GRU layers with progressively decreasing hidden dimensions (256, 124, and 64 neurons, respectively), combined with `tanh` activation functions and a dropout rate of 0.10, consistently achieved

superior stability across folds. Its architecture is therefore selected as the final model configuration and is visualized in Figure 3.

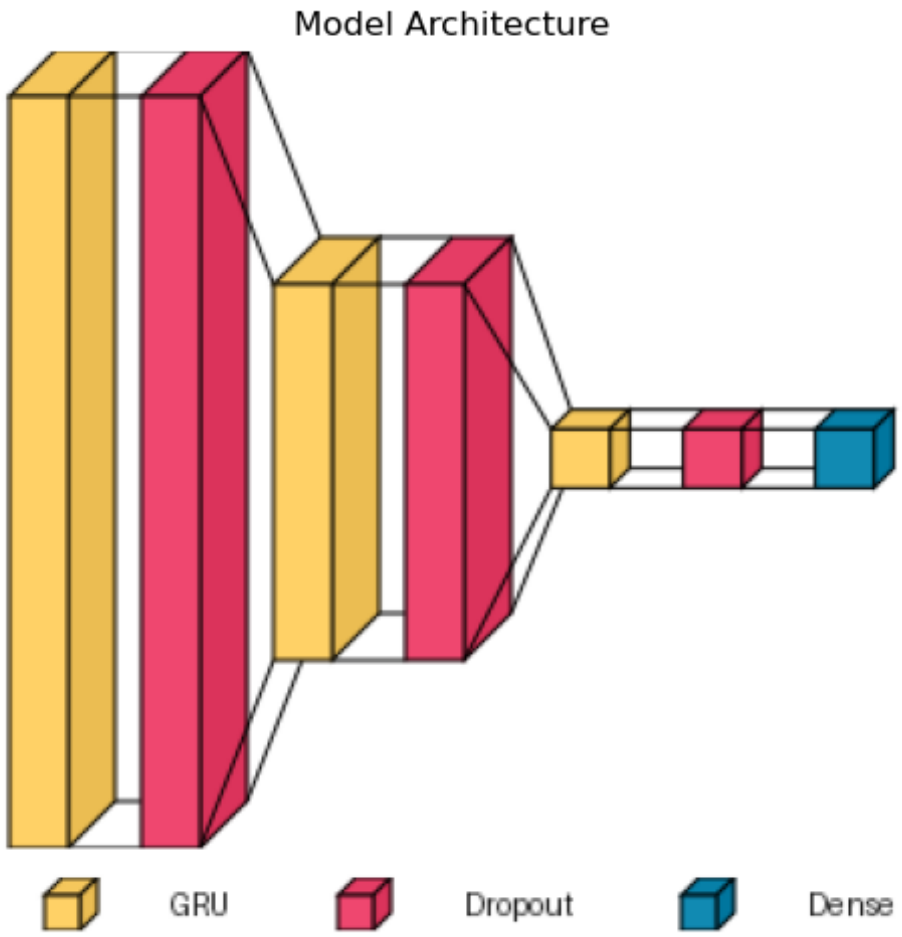

Figure 3: Architecture of the final selected GRU model, consisting of three stacked layers with 256, 124, and 64 units, `tanh` activations, and 0.10 dropout.

3.2.4 Model Retraining and Decision Threshold Optimization

After identifying Model 1 (the three-layer pyramidal GRU) as the most robust and accurate candidate, the architecture was reconstructed and retrained. For this final stage, the entire training dataset was used for fitting, while the full validation dataset was reserved for evaluation. Unlike the earlier cross-validation procedure, which aimed to assess stability and gen-

eralization, this retraining step was designed to optimize the model on the maximum amount of available data while preserving a held-out validation set for unbiased performance measurement.

Once training was complete, an additional optimization step was carried out to determine the most suitable decision threshold for the binary classification of switchbacks. Neural networks trained with a sigmoid activation in the output layer naturally produce probabilities in the range $[0, 1]$, which must be converted into class labels by applying a threshold z. The default choice of $z = 0.5$ often fails under severe class imbalance, as in this case, where switchbacks constituted only $\sim 2.7\%$ of the data (see Section 3.4.2).

To address this, the F_1-score was computed as a function of z. Specifically, for $z \in [0, 1]$ with increments of 0.02, predictions were generated for both the training and validation sets. For each z, labels were assigned as:

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{p} \geq z, \\ 0 & \text{otherwise,} \end{cases}$$

where $\hat{p}$ denotes the predicted probability of a switchback. The F_1-score was then calculated by comparing $\hat{y}$ against the ground-truth labels. This yielded two threshold-dependent F_1 curves, one for the training set and one for the validation set, as shown in Figure 4.

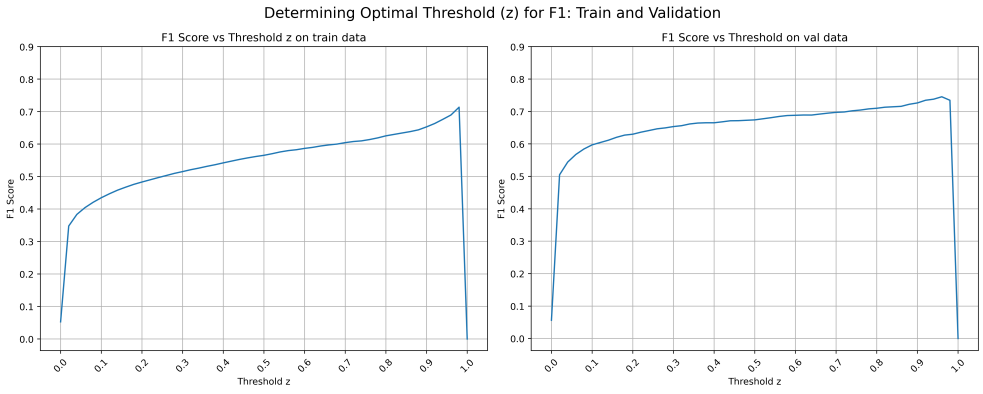

Figure 4: Threshold-dependent F_1 curves for training and validation sets. The vertical dashed line marks the chosen threshold.

The results of this analysis, displayed in Figure 4, show the F_1-score as a function of z for both datasets. Both curves display a steady increase up

to high threshold values, followed by a sharp drop when z approaches 1.0. The optimal value was found at $z = 0.94$, where the validation F_1-score peaked, while the training curve showed consistent performance. This value was therefore adopted as the final classification threshold for all subsequent evaluations of the model.

3.2.5 Model Testing

Once the optimal architecture and decision threshold had been selected (see Section 3.2.4), the model was evaluated on the previously unseen test set to provide an unbiased estimate of its generalization performance. The model outputs were converted into binary predictions using the threshold of 0.94 identified during the validation stage.

Figure 5 shows the resulting confusion matrix, which provides a detailed breakdown of true positives, false positives, true negatives, and false negatives. This visualization allows us to qualitatively assess the model's ability to distinguish between switchback and non-switchback intervals.

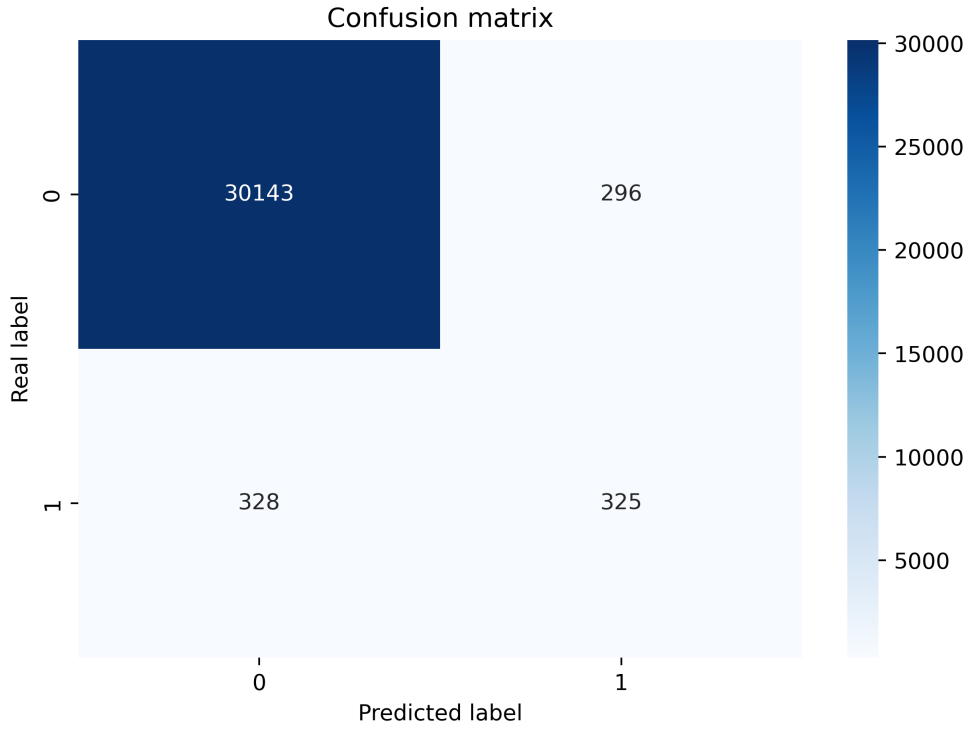

Figure 5: Confusion matrix of the final GRU model evaluated on the test set.

To quantitatively evaluate predictive performance, several standard classification metrics were computed on the test data:

- **Accuracy:** The overall proportion of correctly classified intervals relative to the total number of test samples.
- **Recall (Sensitivity):** The fraction of true switchback events correctly identified, measuring the model's ability to avoid missed detections.
- **Precision:** The fraction of predicted switchbacks that were actually correct, reflecting the reliability of positive predictions.
- **F_1-score:** The harmonic mean of precision and recall, providing a balanced measure particularly suitable for the imbalanced nature of this classification task.

The results obtained on the test data were as follows:

$$\begin{aligned} \text{Accuracy} &= 0.9799 \\ \text{Recall} &= 0.4977 \\ \text{Precision} &= 0.5233 \\ F_1\text{-score} &= 0.5102 \end{aligned}$$

These results indicate that, while the model achieves a very high overall accuracy ($\sim$98%), this figure is largely driven by the overwhelming prevalence of non-switchback intervals. The F_1-score of 0.5102 provides a more balanced perspective, highlighting that the model achieves a moderate compromise between detecting switchbacks and limiting false positives. The recall of 0.4977 shows that approximately half of the true switchbacks are successfully detected, while the precision of 0.5233 indicates that only about half of the predicted switchbacks corresponded to true events.

Taken together, these results confirm that the model is capable of identifying switchbacks at a level substantially better than chance, but also illustrate the inherent difficulty of the task given the rarity and subtlety of switchback signatures in the solar wind data.

4 Future Work

The results obtained in this project indicate that recurrent neural networks are able to detect switchbacks at levels significantly above chance, but the relatively modest F_1-score also highlights the complexity of the problem and the need for further methodological developments. Several promising directions for future research could be:

- **Incorporating additional physical variables:** The present study focused primarily on magnetic field and proton bulk parameters. However, switchbacks are multidimensional phenomena that may also manifest in electron properties, ion composition, plasma beta, or wave activity indicators. Including these additional physical variables could provide richer context for the models and potentially reveal precursors that are not evident in magnetic field data alone.
- **Exploring alternative architectures:** Beyond standard LSTMs and GRUs, more sophisticated architectures could be tested. Convolutional recurrent networks, which combine convolutional layers for local feature extraction with recurrent layers for temporal integration, might be particularly well suited to capturing both short-term fluctuations and longer-term dependencies. Similarly, attention-based architectures or Transformer variants could provide complementary approaches by learning explicit dependencies without recurrence.
- **Addressing class imbalance with synthetic data generation:** Class imbalance remains a key limitation, with switchbacks representing only $\sim 2.7\%$ of the dataset. While class-weighted loss functions were used here, more advanced strategies could improve learning. Generative models such as Generative Adversarial Networks (GANs) could be employed to synthesize realistic switchback-like sequences, augmenting the minority class and reducing bias toward non-switchback intervals.
- **Operational testing and transferability:** Finally, it would be valuable to evaluate these models using other time-ranges when the PSP satellite was at its perihelion and check if similar results are obtained.

In summary, progress in switchback detection will likely depend on combining richer input features, modern sequence modeling architectures, and improved handling of data imbalance. These avenues provide a clear path toward building more accurate and reliable predictive models in future work.

References

[1] Dudok de Wit, T., Krasnoselskikh, V. V., Bale, S. D., Bonnell, J. W., Bowen, T. A., Chen, C. H. K., Froment, C., Goetz, K., Harvey, P. R., Jagarlamudi, V. K., et al. (2020).

[2] Pecora, F., Matthaeus, W. H., Primavera, L., Greco, A., Chhiber, R., Bandyopadhyay, R., and Servidio, S. (2022). Magnetic switchback occurrence rates in the inner heliosphere: Parker Solar Probe and 1 au. *The Astrophysical Journal*, 930(2), 123. doi:10.3847/1538-4357/ac63d2 Switchbacks in the near-Sun magnetic field: Long memory and impact on the turbulence cascade. *The Astrophysical Journal Supplement Series*, 246(2), 39. doi:10.3847/1538-4365/ab5853

[3] Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., and Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, pp. 729-735. doi:10.1109/ICSSIT48917.2020.9214160

[4] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., and Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, 5(1), 1–30.

[5] Zargar, S. A. (2021). Introduction to Sequence Learning Models: RNN, LSTM, GRU. Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC, USA.

[6] Chen, L., and Cui, J. (2023). TPRNN: A Top-Down Pyramidal Recurrent Neural Network for Time Series Forecasting. *arXiv preprint*.