Unravelling Ion-Scale Coherent Structures in the Solar Wind with Machine Learning

Yufei Yang
Department of Physics
Imperial College London
yy3519@ic.ac.uk

Abstract

This paper presents a machine learning (ML) approach to automate the identification of a specific class of ion-scale coherent structures in the solar wind. Tentatively identified as Alfvénic solitons, these structures are characterized by their distinctive magnetic field profiles and are thought to play a role in driving the solar wind turbulence cascade. Traditional detection methods, such as wavelet transforms and non-Gaussianity analysis, are effective but labor-intensive, particularly when handling vast datasets. Our supervised ML framework, trained on a small curated dataset of 466 manually identified events and enhanced with data augmentation, streamlines this process. Among the tested models, Random Forest (RF) stood out for its practicality, prioritizing precision and returning fewer, more standardized samples, making it ideal for human review. Applied to unseen Parker Solar Probe data within 0.25 AU across three years, RF achieved a high precision of 0.92 at a 0.9 classification threshold and identified approximately 500 new events. This detection scale closely aligns with traditional methods while significantly reducing analysis time. The study demonstrates that combining small, high-quality datasets with ML techniques provides a scalable, efficient solution for detecting coherent structures across decades of spacecraft data, paving the way for future discoveries and accelerating research in space physics.

1 Introduction

The application of machine learning (ML) in the physical sciences has demonstrated significant potential for automating complex data analysis tasks. In solar wind physics, the detection of ion-scale coherent structures is critical for advancing our understanding of energy dissipation and plasma heating processes [Alexandrova et al., 2013, Bruno and Carbone, 2013]. This study focuses on Alfvénic solitons, a rare class of ion-scale structures distinguished by their unique magnetic signatures. Recent observations from the Solar Orbiter (SO) and Parker Solar Probe (PSP) have resolved their plasma and magnetic profiles, suggesting they may propagate sunward and contribute to the turbulent cascade. These findings underscore the need to expand their dataset for a more comprehensive understanding of their physical nature. Traditional detection methods, such as wavelet transforms and non-Gaussianity analysis, have proven effective but are time-consuming and labor-intensive, particularly when processing large solar wind datasets. This challenge is compounded in the inner heliosphere, where heightened turbulence complicates detection. To address these issues, we developed a supervised ML classifier to automate the detection process, enabling efficient identification of events on a yearly scale.

Trained on an initial dataset of 466 manually identified events, we evaluated several ML models, including traditional algorithms like Random Forest (RF) [Breiman, 2001] and neural network-based models such as Convolutional Neural Networks (CNNs) [LeCun et al., 1998], Long Short-Term

Machine Learning and the Physical Sciences Workshop, NeurIPS 2024.

Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997], and transformer-based architectures (transformer)[Vaswani et al., 2017]. Among the models tested, although RF did not achieve the highest evaluation metrics across all models, its conservative nature – returning fewer, more standardized samples – made it particularly suited for human review. This inherent bias towards the clearest events, combined with its highest precision as determined through sampling-based evaluation, aligns well with our goal of curating a reliable dataset of well-defined events. When applied to unseen PSP data, RF effectively identified new events in the inner heliosphere, expanding the dataset to nearly 1,000 events. This study makes two key contributions: (1) it streamlines the detection process, significantly expanding the catalog of target ion-scale structures; and (2) it demonstrates the efficacy of ML in identifying small-scale solar wind structures, providing a scalable framework for future studies of similar phenomena in space physics.

2 Background and Related Work

Alfvénic solitons, the ion-scale coherent structures targeted in this study, were first reported by Rees et al. [2006] using Ulysses magnetometer data. These rare solitary structures exhibit sharp magnetic field enhancements and bipolar rotations in the field components, typically governed by ion inertial scales. Minimum Variance Analysis (MVA) reveals their characteristic banana-shaped hodograms. Our recent observations during the February 2022 PSP-SO alignment confirmed their presence across various heliocentric distances, suggesting that these solitons could propagate sunward, contribute to the decay of anti-sunward Alfvénic fluctuations, and drive the turbulent cascade. Example events are provided in Appendix A.1. PSP and SO provide state-of-the-art magnetic and plasma data for studying such ion-scale phenomena. PSP, the fastest human-made object, specializes in near-Sun measurements, recording magnetic field fluctuations up to 20 kHz with the FIELDS suite and reaching distances as close as 0.05 AU [Bale et al., 2016]. SO offers complementary heliospheric coverage, capturing magnetic field measurements at resolutions up to 64 Hz [Horbury et al., 2020], and observes a broader range of heliocentric distances. Leveraging data from these missions enables the detection of magnetic events across diverse spatial and plasma conditions.

Coherent structures represent localized energy concentrations that deviate from Gaussian statistics [Burlaga, 1991, Frisch, 1995, Biskamp, 2003, Miao et al., 2011]. Traditional detection methods rely on frequency filtering techniques, such as wavelet transforms [Farge et al., 1992], followed by non-Gaussianity analyses like Partial Variance of Increments (PVI) [Greco et al., 2018]. While effective, these approaches are time-intensive and require substantial human effort, particularly for analyzing large datasets over multiple years. ML offers a promising alternative. In space physics, ML has been applied to specific tasks, including solar wind classification using k-nearest neighbors [Camporeale et al., 2017], Gaussian Processes [Li et al., 2020], and unsupervised clustering via k-means [Roberts et al., 2020]. However, ML applications for identifying solar wind coherent structures remain underexplored. The closest effort, by Fordin et al. [2023], employed a 1D-CNN to classify circularly polarized waves but focused on broader wave modes rather than specific ion-scale structures like Alfvénic solitons. Furthermore, their study evaluated only CNNs, whereas this work systematically explores a broader range of ML models to identify the most suitable approach, aiming to optimize detection accuracy and robustness.

3 Methods and Results

The workflow involves three key stages: dataset creation, model training and validation, and evaluation on unseen data. With only a limited number of clear events available initially, direct ML classifier development was impractical. Thus, building a robust, high-quality dataset became a critical first step. This involved addressing two key challenges: determining the dataset size necessary for reliable model performance and ensuring its generalizability. Although the methods adopted here proved effective, alternative strategies could be explored in future studies to further refine the approach.

3.1 Dataset Creation

Data Selection: The training data were sourced from SO burst-mode magnetic field data [Horbury et al., 2020] spanning 2020 to 2023, and PSP magnetic field data [Bale et al., 2016] within 0.4 AU from 2020 and 2022. Unseen data for testing comprised PSP encounter data (<0.25 AU) from

2019, 2021, and 2023. These selections allowed the model to be tested in highly turbulent magnetic environments with elevated background fields. Data were accessed via the SunPy library [Mumford et al., 2015]. Testing data were exclusively from PSP because all available SO burst-mode data were used for training. Events from SO, typically observed at larger heliocentric distances with smaller background fields, are clearer and easier to identify manually, making them ideal for building a high-quality training dataset. Future work could extend the model to newly acquired SO observations to evaluate its ability to detect new events and validate its performance across other datasets. **Manual Identification:** Potential ion-scale coherent structures were identified by filtering fractional magnetic field fluctuations ($\delta B / \langle |B| \rangle$) around ion scales using wavelet transforms. A Mexican-Hat wavelet was chosen as the mother function for its shape similarity to the target structure's field enhancement. For SO data, a 1.0-second timescale was used, while PSP employed a 0.5-second scale to account for shorter-duration events near the Sun. A stringent 99.99th percentile threshold was applied to the wavelet results from each day of magnetic field data, reducing the number of timestamps requiring manual inspection to fewer than 20 per day. Each 30-second window containing potential events was evaluated using magnetic field profiles ($B_R, B_T, B_N, |B|$), MVA profiles ($B_{\text{min}}, B_{\text{int}}, B_{\text{max}}$), and hodogram plots ($B_{\text{int}}$ vs. B_{max}) (see Appendix A.3). The selection criteria included: (1) significant $|B|$ enhancements, (2) clear bipolar rotations in at least one RTN or MVA component with associated variations in others, and (3) distinct hodogram patterns resembling banana or quasi-circular shapes. This process yielded 157 events from SO and 309 from PSP, resulting in a combined dataset of 466 events.

Feature Selection: Magnetic field features were included (B_R, B_T, B_N, and $|B|$). To capture subtle patterns not evident in raw data, additional features derived from MVA, B_{min}, B_{int}, and B_{max}, were incorporated. These features reflect variations along the principal axes in the MVA frame, where the characteristics of ion-scale coherent structures are often more distinct. Velocity and electric field data, though potentially relevant, were excluded due to inconsistent availability and lower cadence. Additionally, attempts to incorporate hodograms alongside time series data for training did not yield significant improvements in model performance. Future studies could revisit these approaches to evaluate their utility in enhancing ML-based detection. **Time Window Selection:** A 4-second time window was chosen to capture events ranging from 0.1 to 1 second. This duration was not arbitrary but based on initial experiments testing window sizes between 3 and 5 seconds. The 4-second window consistently demonstrated optimal performance on training and validation datasets, effectively balancing event coverage and model accuracy.Each time window was segmented into 250 data points, determined by the lowest sampling rate across the dataset to ensure uniformity. Downsampling was applied where necessary to preserve the event's essential characteristics while avoiding artifacts introduced by upsampling. This approach ensured that the temporal resolution remained adequate for detecting the targeted ion-scale features. **Data Augmentation:** To enhance model robustness, two augmentation techniques were applied to the positive samples: space reversal (parity) and time reversal. These transformations were chosen because they preserve the pseudovector properties of the magnetic field, ensuring physical consistency (Appendix A.4). Space reversal flips the sign of all magnetic field components, simulating the mirror image of the field structure. Time reversal reverses the temporal sequence of the magnetic field profile while maintaining the spatial relationships between the components. Together, these augmentations effectively doubled the size of the positive sample set, allowing the model to better generalize to different orientations and temporal patterns of ion-scale structures. **True and False Sample Creation:** Given the limited availability of true samples, the above data augmentation was applied to produce two extra folds of true samples. Additional true samples were generated using a high overlap ratio of 0.9 on either side of the original event window. False samples were generated by selecting time windows adjacent to each true event with overlapping ratios of 0.0 and 0.2, producing four false samples per event. To ensure a broader representation of non-events, randomly selected time windows from unrelated regions of the dataset were also included, resulting in eight false samples per event. These false samples were then randomly down-sampled to match the number of true samples to prevent class imbalance. **Data Combination and Normalization:** True and false samples were combined, shuffled, and normalized with a mean of zero and variance of one. Dataset integrity was verified through label balance and visual checks. Each sample now included seven features ($B_R, B_T, B_N, |B|, B_{\text{min}}, B_{\text{int}}, B_{\text{max}}$) and 250 data points, with examples shown in Appendix A.5. Validation results confirmed the merged dataset (4,194 samples from 466 events) outperformed isolated datasets. (Appendix A.6)

3.2 Model Selection, Training, and Validation

Algorithm Selection: A range of machine learning models were evaluated for multivariate time series classification. These included traditional models, such as Random Forest (RF) [Breiman, 2001], XGBoost (XGB) [Chen and Guestrin, 2016], and Support Vector Machines (SVM-C) [Cortes and Vapnik, 1995], as well as advanced architectures like Deep Neural Networks (DNN) [LeCun et al., 2015], Long Short-Term Memory networks (LSTM) [Hochreiter and Schmidhuber, 1997], Convolutional Neural Networks (CNN) [LeCun et al., 1998], and Temporal Convolutional Networks (TCN) [Lea et al., 2017]. Transformer-based architectures [Vaswani et al., 2017] were also explored, along with the Multivariate LSTM Fully Convolutional Network (MLSTM-FCN) and its attention-augmented variant (MALSTM-FCN) [Karim et al., 2019], which are well-suited for capturing both spatial and temporal dependencies. **Model Training and Optimization:** The dataset was split into 80% for training and 20% for a held-out test set, ensuring the test set was exclusively reserved for final evaluation. The training data were further stratified to preserve the class distribution of positive and negative samples. To optimize hyperparameters and prevent overfitting, five-fold stratified cross-validation was performed on the training data. This ensured that each fold maintained the original class proportions and provided robust validation across all splits. Traditional models (e.g., RF, XGB, SVM-C) required flattening the multivariate time series data into a single-dimensional input format. Hyperparameter tuning followed a two-stage approach: first, a randomized search explored a broad parameter space to identify promising configurations, followed by a grid search for fine-tuning. Neural network models (e.g., CNNs, LSTMs) were optimized using systematic searches over common hyperparameters, such as the number of convolutional filters, hidden units, and learning rates, balancing computational efficiency with performance. To mitigate overfitting, early-stopping and adaptive learning rate reduction were applied during training. **Performance Evaluation:** After determining optimal hyperparameters, final models were retrained on the full 80% training set and evaluated on the 20% held-out test set. Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and AUC scores for both ROC and PR curves [Vujović et al., 2021]. Detailed hyperparameter configurations and evaluation results are provided in Appendix A.7 and A.8, respectively.

3.3 Application to Unseen Data

For deployment, models were retrained on the entire labeled dataset to maximize exposure to data variability and enhance generalization. Unseen PSP data were preprocessed identically to the training data. This included wavelet transformation, segmentation into 30-second windows, normalization, and reshaping to match the machine learning input format. Non-overlapping windows were used to minimize redundancy, though this choice may increase false negatives by potentially excluding partial events. Optimal classification thresholds were initially derived from the held-out test set, considering metrics such as precision, ROC-AUC, and F1 score, as detailed in A.8. However, thresholds optimized for precision often yielded too few positive predictions, while those favoring ROC-AUC or F1 score resulted in excessive predictions, overwhelming human review capacity. To balance these considerations, a threshold of 0.9 was selected for simplicity and practicality. At this threshold, we observed that the RF model consistently returned fewer positive samples, yielding a manageable scale of just a few hundred events. This scale is well-suited for human review and aligns closely with event counts from previous manual identification methods. While this conservative approach may exclude ambiguous events, it ensures the identification of clear and reliable examples, supporting our goal of curating a high-quality dataset. To evaluate model precision, a proportional sampling strategy was employed. Ten percent of positive predictions at the 0.9 threshold were randomly selected for manual labeling, and precision was calculated from these samples. The standard error of precision was estimated using a Poisson distribution, appropriate for rare and independent events (Appendix A.9). RF achieved the highest precision under this evaluation method, with RF and XGB emerging as the top-performing models. RF's bagging approach and XGB's boosting approach both underscore the robustness of ensemble methods [Ganaie et al., 2022]. Future research could explore hybrid ensemble techniques, such as majority voting, to integrate strengths from multiple models. Additionally, the potential of neural networks like CNNs and LSTMs, which demonstrated strong F1 scores and ROC-AUC metrics during evaluation, could be revisited to enhance generalizability and capture a broader range of event types. Manual review of RF-identified samples at the 0.9 threshold yielded true positive counts of 67, 229, and 212 for 2019, 2021, and 2023, respectively. These corresponded to precisions of 0.96, 0.91, and 0.90, with an overall precision of 0.92 across all years. Approximately

500 new events were identified, significantly improving detection efficiency compared to traditional methods.

Model	Precision	Uncertainty
RF	0.89	0.12
XGB	0.68	0.03
SVC	0.47	0.03
DNN	0.38	0.02
LSTM	0.67	0.02
CNN	0.54	0.03
TCN	0.66	0.03
Transformer	0.53	0.02
MLSTM_FCN	0.43	0.04
MALSTM_FCN	0.42	0.02

Table 1: Precision and uncertainty for various ML models evaluated on unseen PSP data. Precision was calculated using a proportional sampling approach at a 0.9 threshold, with 10% of positive predictions manually reviewed. The standard error was estimated using a Poisson distribution, assuming rare, independent events with variance equal to the mean number of true positives. The top-performing model, RF, is highlighted in bold for its superior precision evaluated here.

4 Discussion and Limitation

This study demonstrates the potential of ML to automate the detection of ion-scale coherent structures in the solar wind, significantly reducing manual effort while maintaining high precision. Among the tested models, RF emerged as a strong performer, showcasing robust generalizability beyond the training dataset and highlighting the value of combining small, high-quality datasets with ML techniques. Additional feature importance analysis from the RF model identified $|B|$, B_R, B_{int}, and B_{max} as the most significant predictors. These features, which capture key aspects of magnetic field strength and rotational characteristics, align well with theoretical models of Alfvénic structures, reinforcing the physical relevance of the model's predictions.

While this study acknowledges several limitations regarding biases in dataset curation and event selection, these choices were made with practical and scientific considerations in mind. Specifically, the manual identification of events may over-represent well-defined structures while under-representing subtler or more complex ones, introducing potential bias in the dataset. Furthermore, the exclusive use of positive samples for validation precluded direct measurement of false negative rates, potentially favoring the most easily recognizable events. Despite these limitations, these decisions were informed by the need to create a clear, consistent, and high-quality dataset. Addressing these challenges in future work could involve the use of unsupervised or semi-automated labeling techniques to expand event diversity and enhance model recall. The workflow also relied on data from SO and PSP, with unseen data exclusively drawn from PSP encounters. Although the RF model demonstrated strong generalizability across PSP data, its performance under varying heliospheric conditions remains untested. Incorporating newly acquired SO observations or data from other missions could validate its broader applicability and further enhance the diversity of detected events. The decision to prioritize a model, such as RF, that returns the fewest positive samples for manual verification significantly reduced the time and labor required for event validation. While this approach may introduce a bias toward detecting the clearest and most easily defined events, it aligns with the primary objective of establishing a foundational understanding of these rare and underexplored phenomena. By focusing on the most standard and well-defined events, this work built a reliable initial dataset for Alfvénic solitons. Future research will seek to address these biases by expanding event diversity and exploring detection techniques that capture a broader spectrum of these structures, paving the way for a deeper understanding of their role in solar wind turbulence and energy dissipation.

Acknowledgments

The author gratefully acknowledges the data teams from Solar Orbiter and Parker Solar Probe for providing the high-resolution magnetic field measurements that were fundamental to this research. Special thanks are extended to Professor Timothy S. Horbury from Imperial College London for his innovative project vision and dedicated supervision throughout the study. While the full dataset cannot be released at this stage due to ongoing research, additional data examples are provided in Appendix A.3 and A.10 to support transparency and to offer further insights into the study.

References

Olga Alexandrova, Christopher HK Chen, Luca Sorriso-Valvo, Timothy S Horbury, and Stuart D Bale. Solar wind turbulence and the role of ion instabilities. *Space Science Reviews*, 178(2):101–139, 2013.

SD Bale, K Goetz, PR Harvey, et al. The fields instrument suite for solar probe plus. *Space Science Reviews*, 204(1-4):49–82, 2016. doi: 10.1007/s11214-016-0244-5.

Dieter Biskamp. *Magnetohydrodynamic turbulence*. Cambridge University Press, 2003.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Roberto Bruno and Vincenzo Carbone. The solar wind as a turbulence laboratory. *Living Reviews in Solar Physics*, 10(1):2, 2013.

Leonard F Burlaga. Intermittent turbulence in the solar wind. *Journal of Geophysical Research: Space Physics*, 96(A4):5847–5851, 1991.

Enrico Camporeale, Algo Carè, and Joseph E Borovsky. Classification of solar wind with machine learning. *Journal of Geophysical Research: Space Physics*, 122(11):10–910, 2017.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Marie Farge et al. Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, 24(1):395–458, 1992.

Samuel Fordin, Michael Shay, Lynn B Wilson III, Bennett Maruca, and Barbara J Thompson. A machine learning–based approach to time-series wave identification in the solar wind. *The Astrophysical Journal*, 949(2):40, 2023.

Uriel Frisch. *Turbulence: the legacy of AN Kolmogorov*. Cambridge university press, 1995.

Mudasir A Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022.

A Greco, WH Matthaeus, S Perri, KT Osman, S Servidio, M Wan, and P Dmitruk. Partial variance of increments method in solar wind observations and plasma simulations. *Space Science Reviews*, 214:1–27, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

TS Horbury, H O'Brien, I Carrasco Blazquez, M Bendyk, P Brown, R Hudson, V Evans, et al. The solar orbiter magnetometer. *Astronomy & Astrophysics*, 642:A9, 2020. doi: 10.1051/0004-6361/201937257.

Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Multivariate lstm-fcns for time series classification. *Neural Networks*, 116:237–245, 2019.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Hui Li, Chi Wang, Cui Tu, and Fei Xu. Machine learning approach for solar wind categorization. *Earth and Space Science*, 7(5):e2019EA000997, 2020.

Wengang Miao, Baolin Peng, Yinghui Li, and Li Zhou. Intermittency in magnetohydrodynamic turbulence: An overview. *Physics of Plasmas*, 18(11):112306, 2011.

Stuart J Mumford, Steven Christe, David Pérez-Suárez, Jack Ireland, Albert Y Shih, Andrew R Inglis, Simon Liedtke, Russell J Hewett, Florian Mayer, Keith Hughitt, et al. Sunpy—python for solar physics. *Computational Science & Discovery*, 8(1):014009, 2015.

D Perrone, O Alexandrova, A Mangeney, M Maksimovic, C Lacombe, V Rakoto, JC Kasper, and D Jovanovic. Compressive coherent structures at ion scales in the slow solar wind. *The Astrophysical Journal*, 826(2):196, 2016.

Sebastian Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

A Rees, A Balogh, and TS Horbury. Small-scale solitary wave pulses observed by the ulysses magnetic field experiment. *Journal of Geophysical Research: Space Physics*, 111(A10), 2006.

D Aaron Roberts, Homa Karimabadi, Tamara Sipes, Yuan-Kuen Ko, and Susan Lepri. Objectively determining states of the solar wind using machine learning. *The Astrophysical Journal*, 889(2): 153, 2020.

Owen Wyn Roberts, Xing Li, and Bo Li. Kinetic plasma turbulence in the fast solar wind measured by cluster. *The Astrophysical Journal*, 769(1):58, 2013.

Bengt UO Sonnerup. Minimum and maximum variance analysis. *Analysis methods for multi-spacecraft data*, 1:185, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Ž Vujović et al. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606, 2021.

A Appendix

A.1 Alfvén Solitons: Physics Background and Examples

Figure 1 presents an Alfvén soliton event observed in Ulysses data as reported by Rees et al. [2006]. The magnetic field profile in Figure 1(a) exhibits a sharp enhancement, while the accompanying banana-shaped hodogram in Figure 1(b) reveals characteristic magnetic field rotations, hallmark features of Alfvén solitons.

Figure 2 illustrates examples of Alfvén soliton events identified in data from (a, b) Solar Orbiter and (c) Parker Solar Probe. Each panel showcases distinct events, featuring sharp magnetic field enhancements and bipolar signatures, which are defining characteristics of soliton structures in the solar wind.

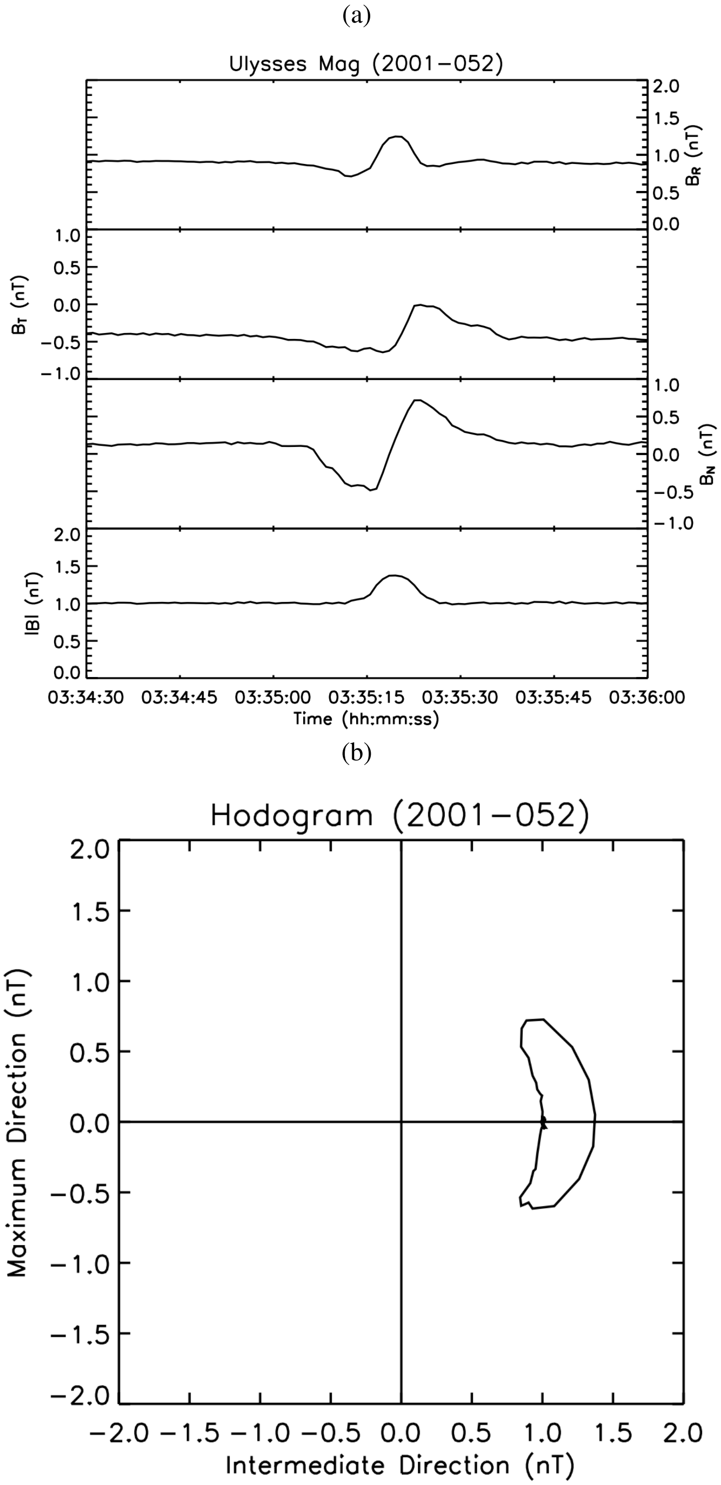
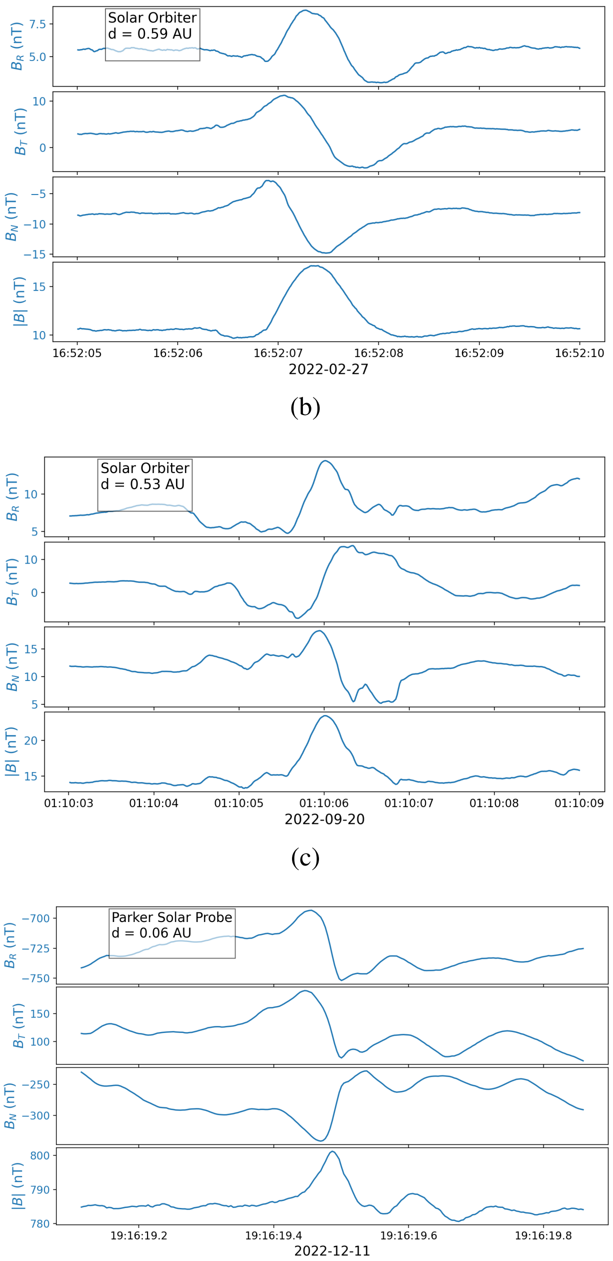


Figure 1: (a) Magnetic field profile and (b) corresponding banana-shaped hodogram of an Alfvén soliton event observed in Ulysses data, as reported by Rees et al. [2006]. The magnetic profile shows a sharp field enhancement, and the hodogram exhibits the characteristic rotation of magnetic field components.

(a)



(b)

(c)

Figure 2: Magnetic field profiles of the solitary events observed by (a), (b), Solar Orbiter, and (c) Parker Solar Probe. Each panel shows a distinct event, characterized by sharp magnetic enhancements and bipolar field signatures indicative of the structure.

A.2 Wavelet Transform

The wavelet transform is a powerful tool for analyzing localized, transient events in time-series data, particularly useful in the context of solar wind turbulence [Roberts et al., 2013, Perrone et al., 2016]. We employed the Mexican Hat wavelet, also known as the Ricker wavelet, which is well-suited for detecting Gaussian-shaped structures with two dips on either side, closely resembling the wavelet's form.

Specifically, the **continuous wavelet transform** was implemented in Python using the `pywt.cwt` function from the `PyWavelets` library. This function applies the wavelet transform to time-series data, capturing fluctuations in magnetic field strength across multiple scales. The transform is mathematically defined as:

$$W_s(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(t)\, \psi^* \left(\frac{t - b}{a} \right) dt \tag{1}$$

where $W_s(a, b)$ represents the wavelet coefficient at scale a and position b, and $\psi(t)$ is the wavelet function. In this case, we used the Mexican Hat wavelet, which is defined as:

$$\psi_{\text{Mexican Hat}}(t) = \frac{1}{\sqrt{2\pi}} \left(1 - t^2\right) e^{-t^2/2} \tag{2}$$

The scale parameter a determines the level of detail, while the translation parameter b shifts the wavelet over the time series to detect localized fluctuations. This wavelet transform was applied to normalized magnetic fluctuations $\delta B / \langle |B| \rangle$ using a 1.0-second scale for Solar Orbiter data and a 0.5-second scale for Parker Solar Probe data, tailored to capture shorter-duration events around ion-scales nearer the Sun.

A.3 Minimum Variance Analysis (MVA) and Hodogram

Minimum Variance Analysis (MVA) is used to transform magnetic field data into a coordinate system aligned with the direction of minimum variance, enhancing the clarity of bipolar rotations within the data. This technique works by calculating the covariance matrix of the magnetic field components and performing an eigenvalue decomposition [Sonnerup, 1998]. The eigenvector associated with the smallest eigenvalue identifies the direction of minimum variance, B_{min}, where the magnetic field varies the least. The other two eigenvectors correspond to the intermediate variance direction B_{int}, and the maximum variance direction B_{max}.

To visualize these results, **hodograms** are generated, graphically representing the changes in the magnetic field's direction and magnitude. In these plots, two components of the magnetic field data, B_{int} and B_{max}, are plotted against each other over a specific time interval. Each point on the hodogram corresponds to the magnetic field vector at a given time.

For an event to be classified as one of the target ion-scale coherent structures, the hodogram should exhibit banana-shaped, elliptical, or nearly circular patterns, forming closed or nearly closed loops with smooth and clean lines. Occasionally, hodograms with other patterns will be accepted if the events satisfy the first two criteria: 1) a significant magnetic field enhancement, and 2) clear bipolar rotations in the RTN or MVA components, accounting for the effects of turbulent solar wind conditions. This signature helps distinguish ion-scale coherent structures from noise or unrelated solar wind features. For each event, the local maxima in the field magnitude $|B|$ were identified and recorded as the event time.

A selection of event examples is presented in Figures 3 and 4.

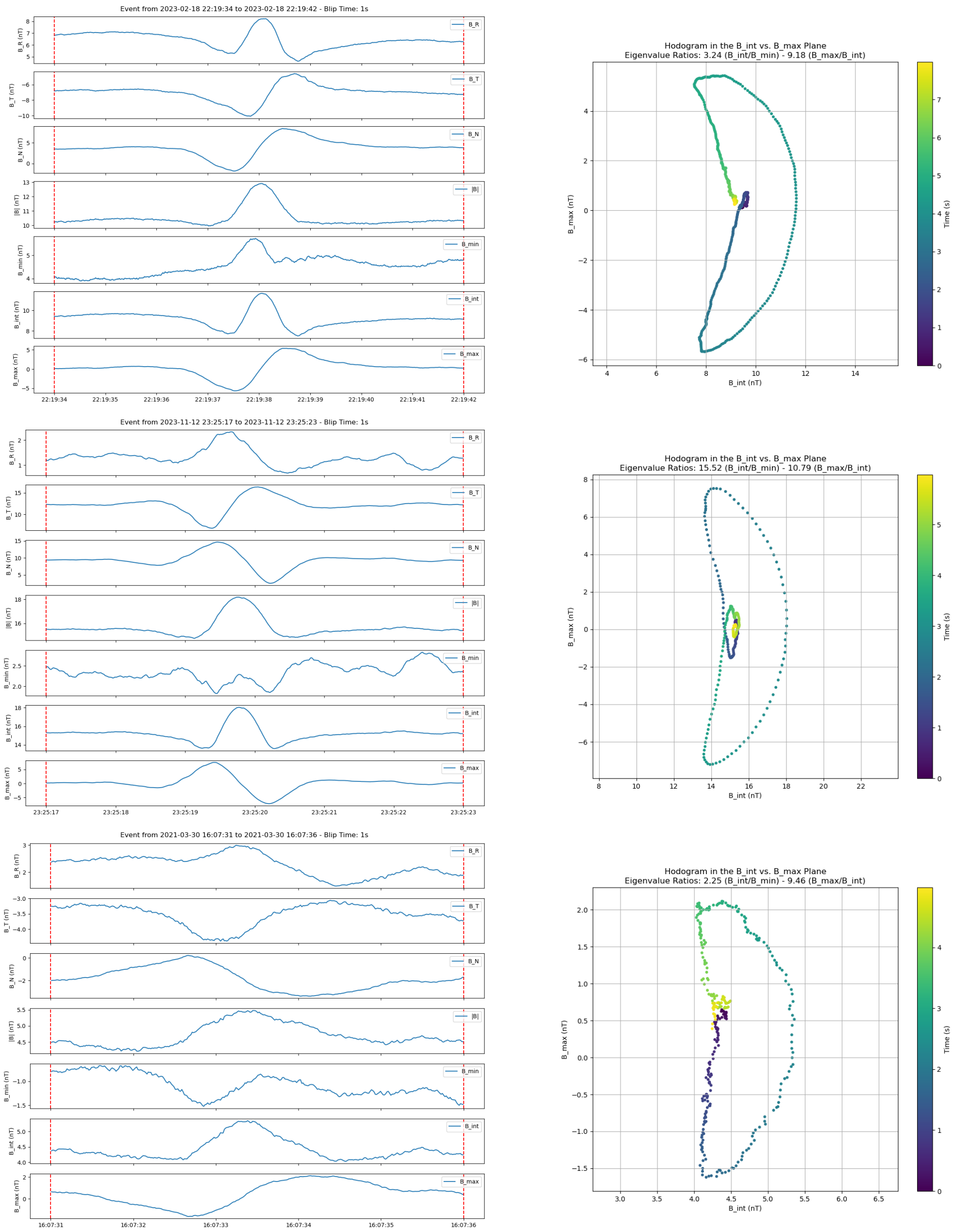


Figure 3: Three example events from the SO dataset identified using the MVA. Each row represents a distinct event. The left panels display the RTN magnetic profiles and the associated MVA profiles over time. The right panels present the corresponding hodograms, plotting B_{int} against B_{max}. The first two events show well-defined banana-shaped hodograms, while the third has a less smooth shape.

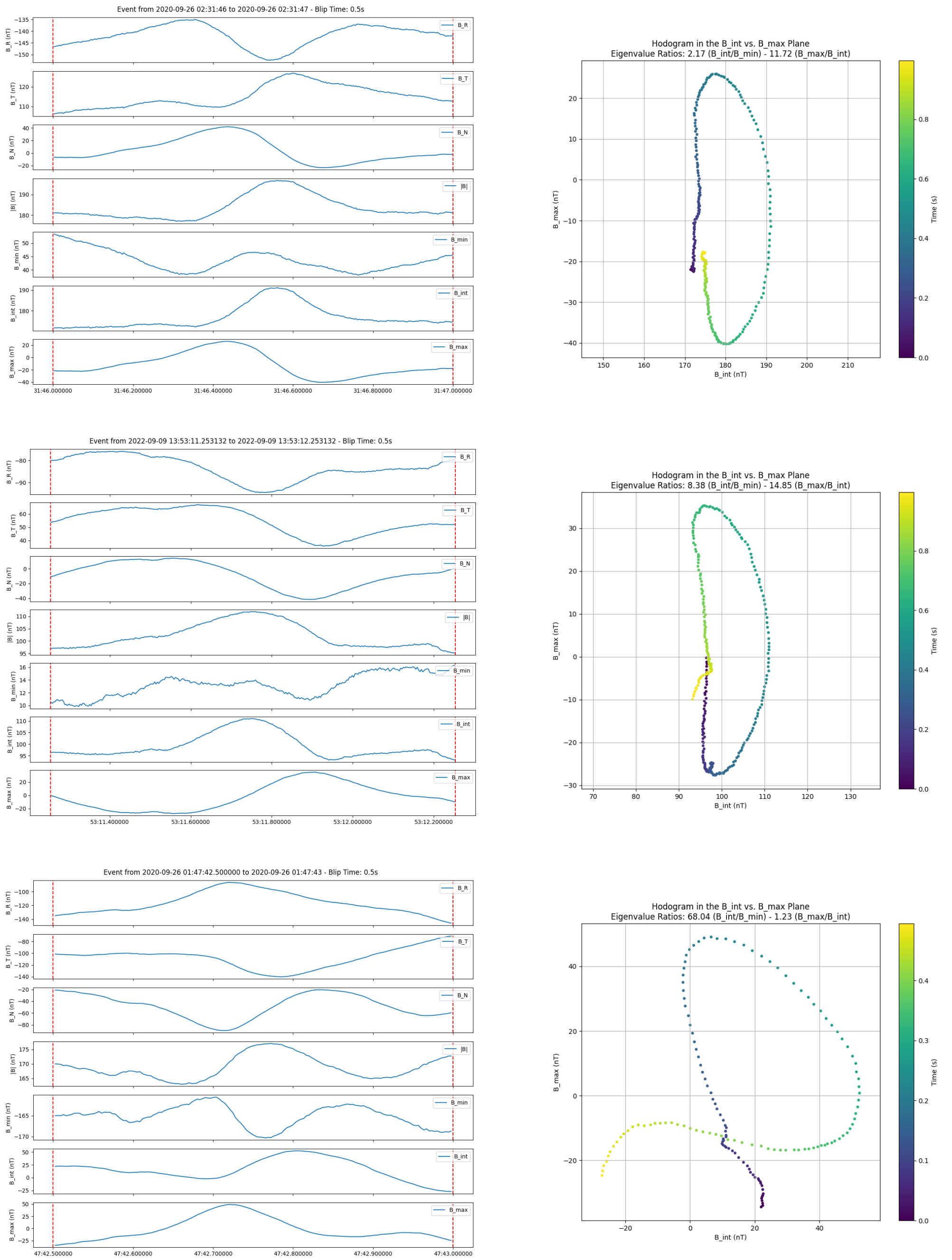


Figure 4: Three example events from the Parker Solar Probe dataset identified using the MVA. Each row represents a distinct event. The left panels display the RTN magnetic profiles and the associated MVA profiles over time. The right panels present the corresponding hodograms, plotting B_{int} against B_{max}. The first two events exhibit well-defined banana-shaped hodograms, while the third shows a more elliptical pattern.

A.4 Data Augmentation: Space and Time Reversal Operators

We applied two key transformations for data augmentation: the **space reversal operator** and the **time reversal operator**, both designed to preserve the fundamental physical properties of the magnetic field while expanding the training dataset.

The **space reversal operator** (parity transformation) inverts the sign of the magnetic field components, ensuring that the magnetic field vectors (B_R, B_T, B_N) and the MVA components (B_{min}, B_{int}, B_{max}) maintain their physical meaning. Mathematically, this transformation is expressed as:

$$\vec{B} \to -\vec{B} \quad \text{or} \quad B'_i = -B_i \quad \text{for} \quad i \in \{R, T, N, \text{min}, \text{int}, \text{max}\} \tag{3}$$

The **time reversal operator** reverses the time sequence while simultaneously flipping the sign of the magnetic field components. This transformation ensures that the physical properties of the data remain consistent across temporal inversions. The time reversal operator is expressed as:

$$\vec{B}(t) \to -\vec{B}(-t) \quad \text{or} \quad B'_i(t) = -B_i(-t) \quad \text{for} \quad i \in \{R, T, N, \text{min}, \text{int}, \text{max}\} \tag{4}$$

Both space and time reversals were applied to the magnetic field components and MVA-derived vectors, effectively doubling the training data while preserving the event's inherent characteristics. Figure 5 illustrates these transformations using synthetic magnetic field data.

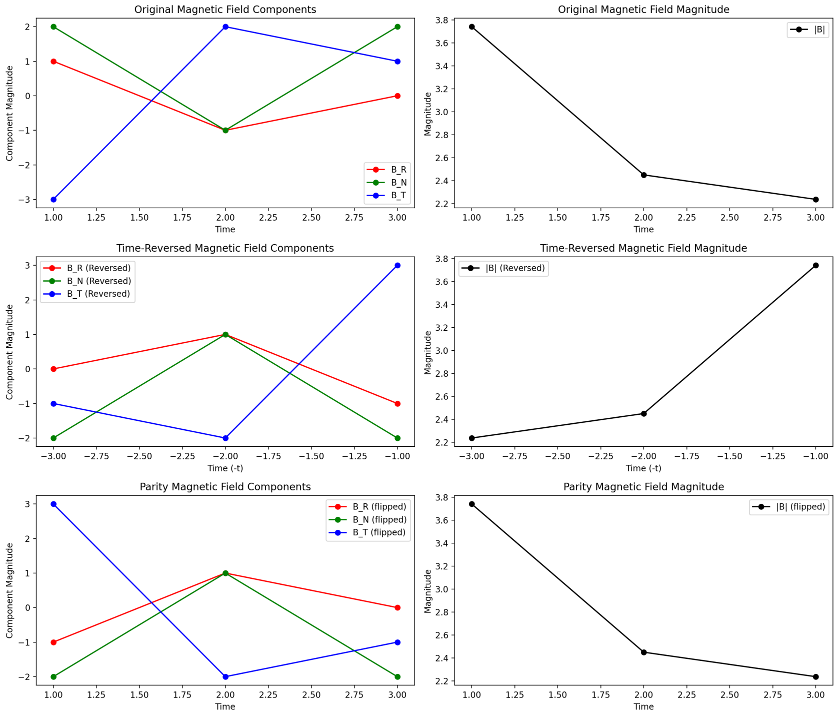


Figure 5: Demonstration of space and time reversal operators applied to synthetic magnetic field data (B_R, B_T, B_N) and the corresponding magnitude $|B|$. The left panels show the RTN components, while the right panels display $|B|$. The first row presents the original data, the second row shows the results after applying time reversal, and the third row illustrates the effects of space reversal.

A.5 ML Dataset Example

Figures 6 and 7 illustrate randomly selected data samples from the final shuffled event datasets for SO and PSP, respectively. These examples showcase the key features used in training the machine learning models.

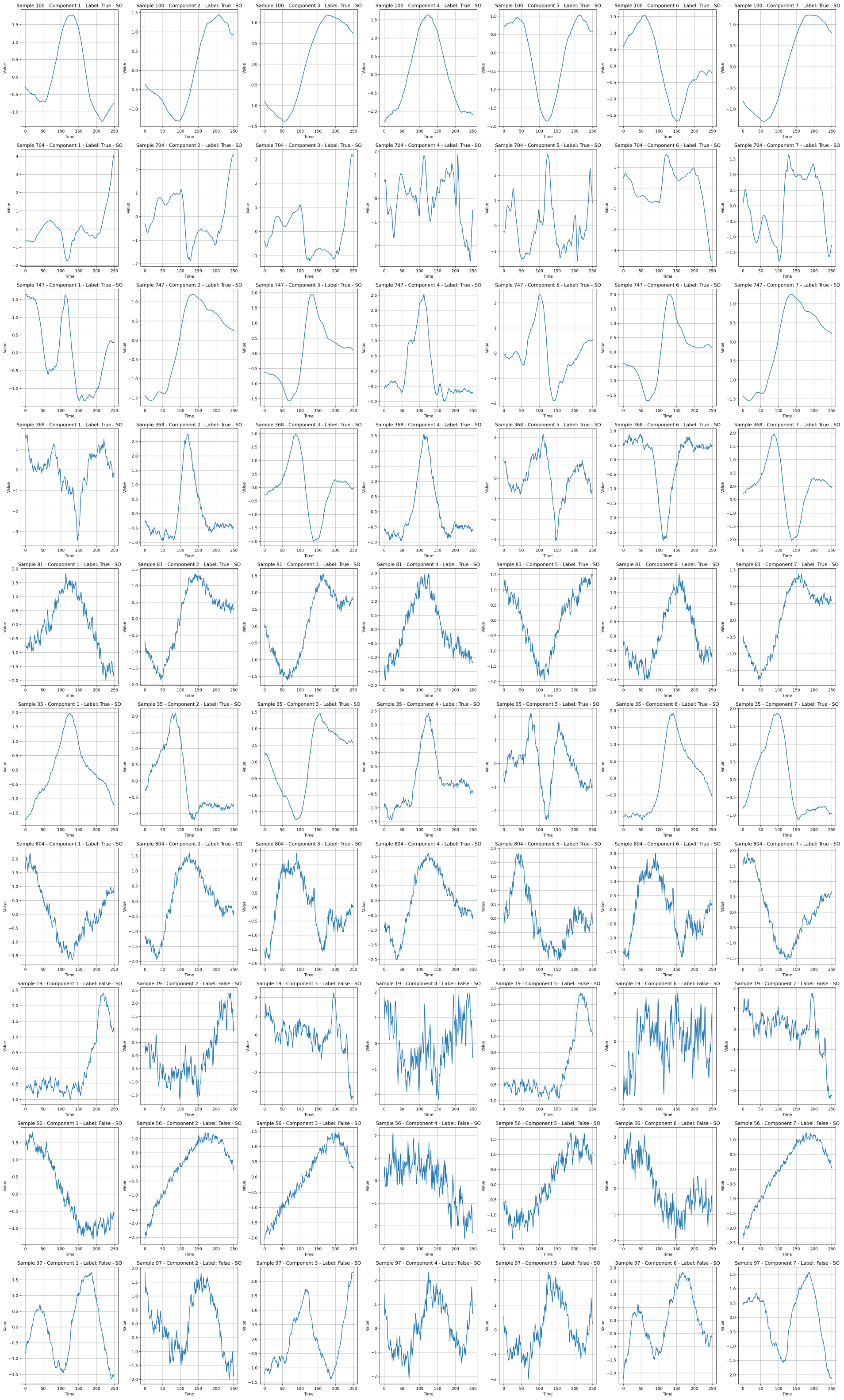


Figure 6: Ten randomly selected samples from the final shuffled SO dataset. Each row corresponds to one sample, with seven sequentially plotted features: magnetic field components B_R, B_T, B_N, the magnitude $|B|$, and MVA components B_{min}, B_{int}, B_{max}. The y-axis represents normalized values, while the x-axis denotes the time steps (0–250) across 250 data points.

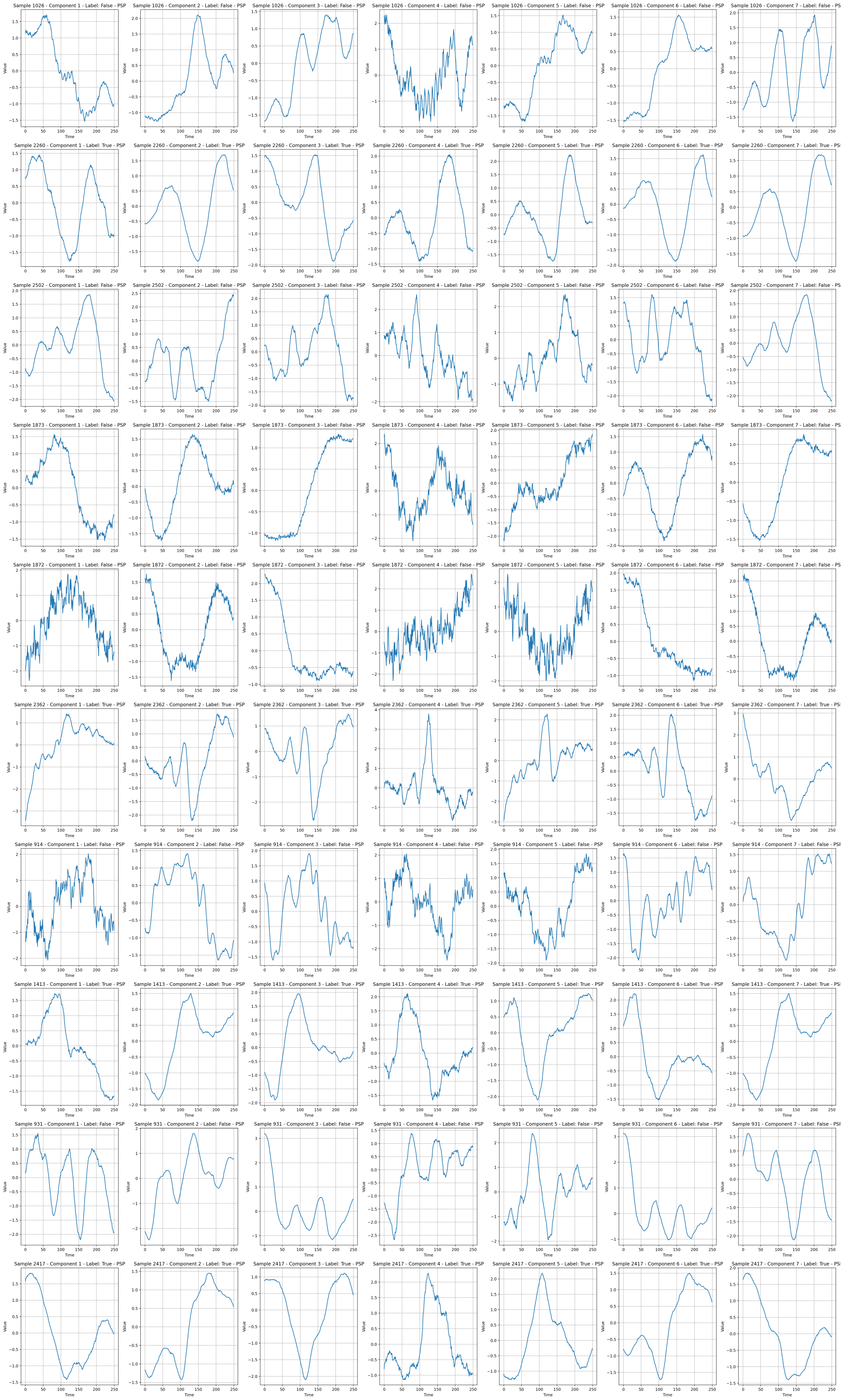


Figure 7: Ten randomly selected samples from the final shuffled PSP dataset. Each row represents one sample, displaying the magnetic field components B_R, B_T, B_N, the magnitude $|B|$, and the MVA components B_{min}, B_{int}, B_{max}. The y-axis represents normalized values, and the x-axis indicates the time steps (0–250) across 250 data points.

A.6 Validation of Merging Datasets

To assess whether merging the SO and PSP datasets would impact model performance, we trained three Random Forest models: `rf_model_so` (trained on SO data), `rf_model_psp` (trained on PSP data), and `rf_model` (trained on the combined dataset). Each model was evaluated on test sets extracted from SO, PSP, and the merged dataset, using accuracy, precision, recall, and F1 score as performance metrics.

Model	Test Set	Accuracy	Precision	Recall	F1 Score
`rf_model`	Full test set	0.921	0.936	0.924	0.930
	SO test set	0.965	0.986	0.947	0.966
	PSP test set	0.993	0.987	1.000	0.993
`rf_model_so`	Full test set	0.814	0.930	0.727	0.816
	SO test set	0.901	0.896	0.920	0.908
	PSP test set	0.749	0.865	0.636	0.733
`rf_model_psp`	Full test set	0.936	0.986	0.899	0.941
	SO test set	0.838	0.933	0.747	0.830
	PSP test set	0.946	0.953	0.947	0.950

Table 2: Performance metrics (accuracy, precision, recall, and F1 score) for three Random Forest models: `rf_model_so` (SO data), `rf_model_psp` (PSP data), and `rf_model` (merged dataset). Each model was evaluated on test sets from SO, PSP, and the merged dataset.

The `rf_model`, trained on the merged dataset, consistently outperformed the other two models, demonstrating enhanced generalization when combining SO and PSP data. By contrast, `rf_model_so` and `rf_model_psp` underperformed on test sets derived from the other dataset, indicating limited cross-dataset generalizability. The merged dataset offered the most robust and reliable performance, supporting its use for further analysis.

A.7 Model Architecture and Hyperparameter

The architecture of each model was carefully designed to maximize predictive accuracy and computational efficiency, implemented using the Tensorflow library, with each of model's architecture and the final hyperparameters briefly specified as below:

- **Random Forest (RF)**: An ensemble method using 500 decision trees to handle high-dimensional data with the Gini criterion.
  - **Hyperparameters**: `criterion='gini'`, `max_depth=24`, `max_features='sqrt'`, `min_samples_leaf=1`, `min_samples_split=10`, `n_estimators=500`.
- **XGBoost (XGB)**: A gradient boosting framework optimized with 500 boosting rounds and decision trees as base learners.
  - **Hyperparameters**: `colsample_bytree=0.75`, `gamma=0.3`, `learning_rate=0.042`, `max_depth=9`, `min_child_weight=1`, `n_estimators=500`, `subsample=0.8`.
- **Support Vector Machine (SVM-C)**: A classification technique using a Radial Basis Function (RBF) kernel to find the optimal hyperplane.
  - **Hyperparameters**: `C=50`, `gamma=0.0005`, `kernel='rbf'`.
- **Deep Neural Networks (DNN)**: A neural network with three hidden layers (224, 192, and 224 neurons) for modeling complex relationships.
  - **Hyperparameters**: `units_first=224`, `dropout_first=0.5`, `num_layers=1`, `units_0=192`, `dropout_0=0.5`, `learning_rate=0.0001`, `units_1=224`, `dropout_1=0.5`, `units_2=160`, `dropout_2=0.2`.
- **Long Short-Term Memory (LSTM)**: Recurrent neural networks with three LSTM layers, designed to capture temporal dependencies in sequential data.
  - **Hyperparameters**: `units=(128, 128, 256)`, `dropout_rate=0.2`, `batch_size=16`, `learning_rate=0.001`.

- **Convolutional Neural Networks (CNN)**: Designed for capturing spatial dependencies in time-series data with three convolutional layers.
  - **Hyperparameters**: `filters=(128, 128, 256)`, `dense_units=(128, 64)`, `dropout_rate=0.4`, `batch_size=16`, `learning_rate=0.001`.
- **Temporal Convolutional Networks (TCN)**: Utilizes dilated convolutions to capture long-range dependencies in sequential data.
  - **Hyperparameters**: `nb_filters=64`, `kernel_size=3`, `nb_stacks=1`, `dropout_rate=0.3`, `dilations=[1, 2, 4, 8, 16]`, `learning_rate=0.001`, `batch_size=16`.
- **Transformer-based Architectures (Transformer)**: Leverages self-attention mechanisms for global dependency capture in sequential data.
  - **Hyperparameters**: `embed_dim=64`, `num_heads=8`, `ff_dim=64`, `rate=0.2`, `learning_rate=0.001`, `batch_size=32`.
- **Multivariate LSTM Fully Convolutional Network (MLSTM-FCN)**: Combines LSTM and fully convolutional networks for spatial and temporal modeling.
  - **Hyperparameters**: `filters1=64`, `filters2=256`, `filters3=128`, `kernel_size1=3`, `kernel_size2=3`, `kernel_size3=3`, `lstm_units=64`, `dense_units=64`, `dropout_rate=0.2`, `learning_rate=0.001`, `batch_size=32`.
- **Multivariate Attention LSTM Fully Convolutional Network (MALSTM-FCN)**: Extends MLSTM-FCN by incorporating attention mechanisms within LSTM layers.
  - **Hyperparameters**: `filters1=128`, `filters2=256`, `filters3=64`, `kernel_size1=3`, `kernel_size2=3`, `kernel_size3=3`, `lstm_units=64`, `dense_units=64`, `dropout_rate=0.2`, `learning_rate=0.001`, `batch_size=16`.

A.8 Model Evaluation on Test Set

Below are commonly used metrics for classification model comparison. While test set performance provides some indication of effectiveness, it's important to recognize that a model's performance on the test set may not fully reflect its actual ability to classify true instances in an unseen dataset. [Raschka, 2018]

Threshold-Based Evaluation Metrics [Vujović et al., 2021]

- **Accuracy:** Proportion of correctly classified instances out of the total instances.

$$\text{Accuracy} = \frac{\text{True Positives + True Negatives}}{\text{Total Observations}} \tag{5}$$

- **Precision:** Ratio of true positive instances to the total predicted positive instances. High precision indicates a low rate of false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}} \tag{6}$$

- **Recall (Sensitivity):** Ratio of true positive instances to the total actual positive instances, reflecting the model's ability to detect positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \tag{7}$$

- **F1 Score:** Harmonic mean of precision and recall, balancing both metrics.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Figure 8 compares the performance metrics across models on the test dataset, evaluated with a classification threshold of 0.5. Bootstrapping with 500 resamples was applied to estimate the error margins for these metrics. Based on the results, CNN and LSTM models performed the best, ranking in the first tier. TCN and Transformer models followed in the second tier, while Random Forest (RF) and XGBoost (XGB) ranked in the third tier.

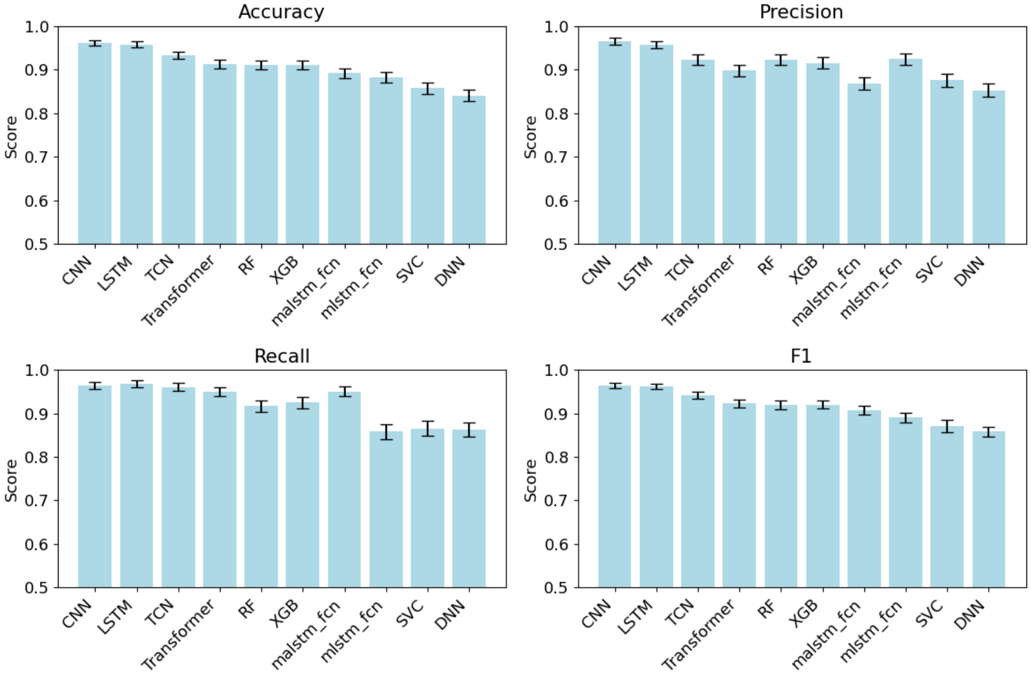


Figure 8: Comparison of model performance based on evaluation metrics (accuracy, precision, recall, and F1 score) on the held-out test dataset using a classification threshold of 0.5 under a bootstrapping process of 500.

Threshold-Free Metrics: ROC Curve, PR Curve, and AUC Scores [Vujović et al., 2021]

While threshold-based metrics are informative, threshold-free metrics such as the ROC curve, PR curve, and their corresponding AUC (Area Under the Curve) scores provide a more comprehensive evaluation of model performance across all possible thresholds. These curves visualize the trade-offs between different metrics and can assist in identifying optimal thresholds.

- **ROC Curve and AUC:** The ROC curve plots the true positive rate against the false positive rate across various thresholds. The AUC represents the model's ability to discriminate between positive and negative classes across all thresholds.

- **PR Curve and AUC:** The PR curve plots precision against recall across different thresholds, particularly useful when dealing with imbalanced datasets. The AUC of the PR curve highlights the balance between precision and recall, emphasizing model performance on positive class predictions.

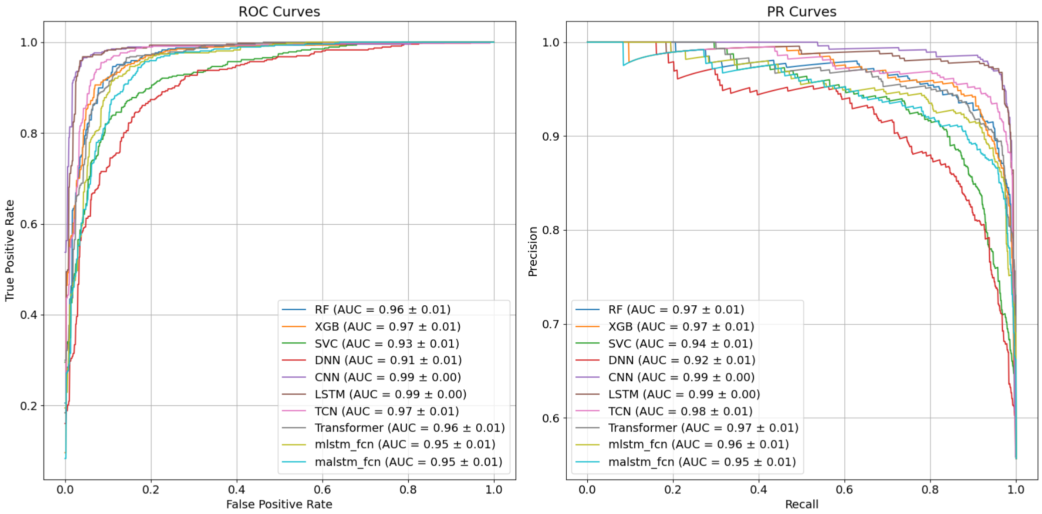


Figure 9: left Panel: ROC curves, Right Panel: PR curves , each with AUC scores labeled on the legend with standard deviations evaluated from bootsrapping with 100 resampling, plotted among models for a direct comparison

Figure 9 presents the ROC and PR curves for each model, with AUC scores indicated in the legend, including standard deviations calculated from 100 bootstrap resamples. Among the models, CNN and LSTM achieved the highest AUC scores, demonstrating superior performance. The TCN and XGB network emerged as a strong second tier, while RF and the Transformer models ranked in the third tier.

Classification Threshold Selection

- **Optimal Threshold from ROC Curve:** This strategy selects the threshold closest to the top-left corner of the ROC curve, achieving a balance between sensitivity and specificity.
- **Threshold Maximizing Precision:** This approach focuses on maximizing precision, which is critical when minimizing false positives is a priority, ensuring the reliability of positive predictions.
- **Threshold Maximizing F1-Score:** This method balances precision and recall to provide a comprehensive evaluation of model performance.

Table 3 shows the optimal thresholds for each model based on different criteria.

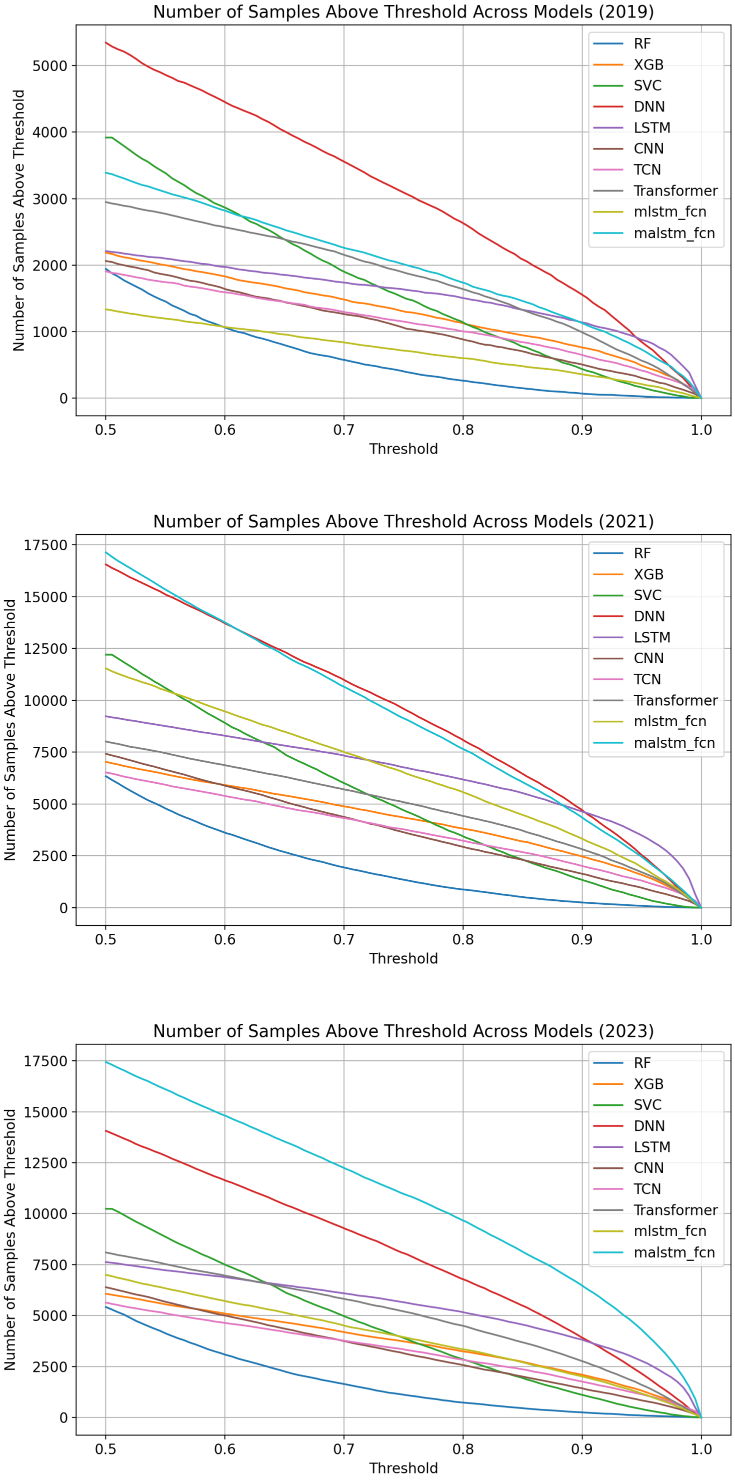
Model	ROC Turning Point Threshold	Optimal F1 Threshold	Optimal Precision Threshold
RF	0.507	0.439	0.980
XGB	0.688	0.688	0.999
SVC	0.485	0.419	0.968
DNN	0.590	0.237	0.998
CNN	0.496	0.496	0.966
LSTM	0.587	0.587	0.998
TCN	0.739	0.628	0.999
Transformer	0.663	0.374	0.995
MLSTM-FCN	0.323	0.201	0.992
MALSTM-FCN	0.817	0.475	0.999

Table 3: Optimal thresholds determined for each model using the ROC turning point, optimal F1, and precision thresholds.

A.9 Application to Unseen Data

1. Number of Samples Above Threshold Across Models

Figure 10 shows the number of samples exceeding the classification threshold for each model. It is evident that the RF model consistently identified the fewest samples for further review, particularly across thresholds ranging from 0.6 to 1.0.



Number of Samples Above Threshold Across Models (2019)

Number of Samples Above Threshold Across Models (2021)

Number of Samples Above Threshold Across Models (2023)

Figure 10: Comparison of the number of samples exceeding the classification threshold across models for unseen data from the PSP in 2019, 2021, and 2023.

2. Precision Uncertainty Evaluation

To estimate the uncertainty in precision, we assume a Poisson distribution, appropriate when true positives are relatively rare and occur independently. The variance of a Poisson-distributed variable is equal to its mean, so the standard deviation of the expected number of true positives, denoted as λ_{TP}, is given by $\sqrt{\lambda_{\text{TP}}}$. To calculate the standard error of the precision, this standard deviation is scaled by the total number of predicted positives N in the dataset, resulting in:

$$\text{Standard Error}_{\text{Precision}} = \frac{\sqrt{\lambda_{\text{TP}}}}{N} \tag{9}$$

Here, $\lambda_{\text{TP}} = \text{Precision} \times N$ represents the expected number of true positives among the predicted positives in the dataset, based on the observed precision in the sample. This approach allows us to express precision as $\frac{\lambda_{\text{TP}}}{N} \pm \frac{\sqrt{\lambda_{\text{TP}}}}{N}$, providing a confidence interval around the estimated precision.

A.10 Additional Dataset Examples

This research culminated in a dataset comprising 974 identified events. Additional examples, including their magnetic profiles and corresponding hodograms, are provided to highlight two key aspects: (1) the rationale for not imposing overly strict event selection criteria, and (2) how these criteria facilitate the discovery of comparable structures by the ML model.

The magnetic structures in the dataset can be categorized into four distinct types based on their hodogram shapes:

- **Banana-Shaped Hodogram:** Characterized by a banana-like curvature (Figure 11).
- **Elliptical-Shaped Hodogram:** Forms an elongated elliptical shape, primarily along the B_{int} axis (Figure 12).
- **Triangle-Like Hodogram:** Resembles a triangular pattern (Figure 13).
- **Irregular Hodogram:** Displays an irregular and less structured pattern (Figure 14).

These categories capture varying levels of loop completion and smoothness, ranging from well-formed structures to more dynamic and complex configurations. Figures in this section illustrate representative examples from each category, showcasing both distinct, well-formed patterns and less defined structures. The upper panels in each figure correspond to events included in the manually curated ML training set, while the lower panels depict events discovered by the Random Forest classifier. This juxtaposition demonstrates the variety and complexity of events captured within the dataset, underscoring the robustness of the ML approach in identifying diverse solar wind structures.

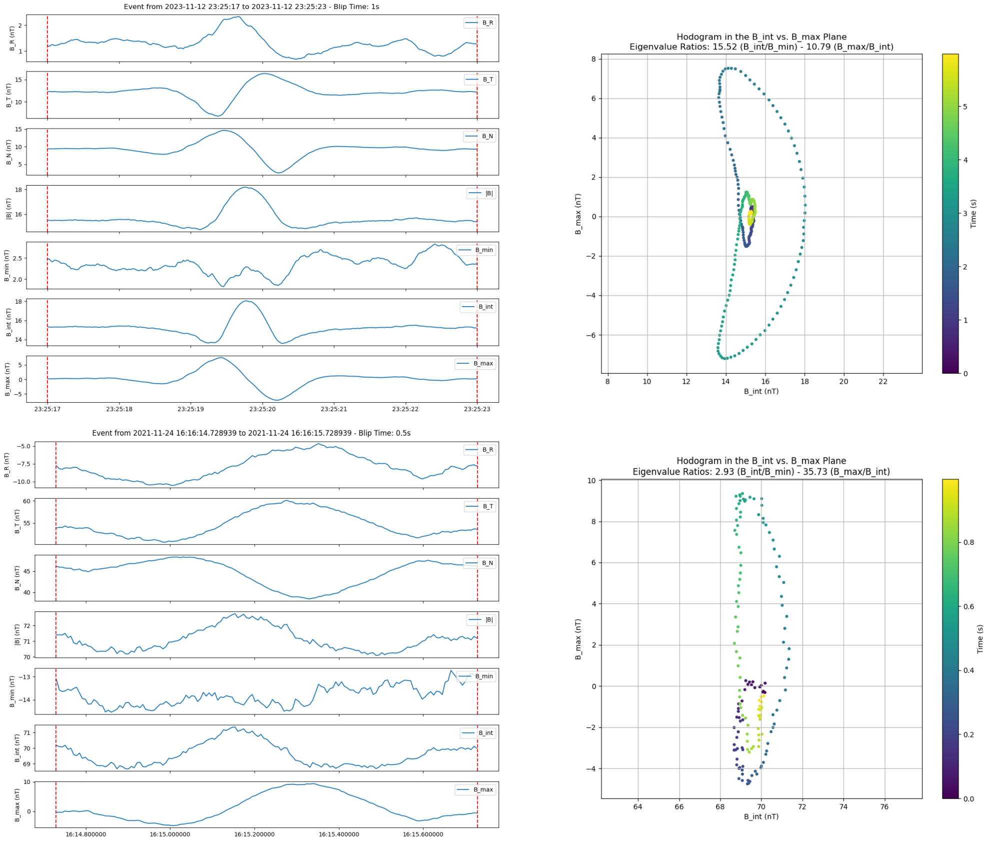


Figure 11: Examples of banana-shaped hodograms. The upper panel shows an event from the 2023 SO dataset, and the lower panel shows a similar event identified by the RF classifier from the unseen 2021 PSP data. Left panels display the RTN magnetic field profiles and the associated MVA profiles over time, while right panels present the corresponding hodograms, plotting B_{int} v.s. B_{max}.

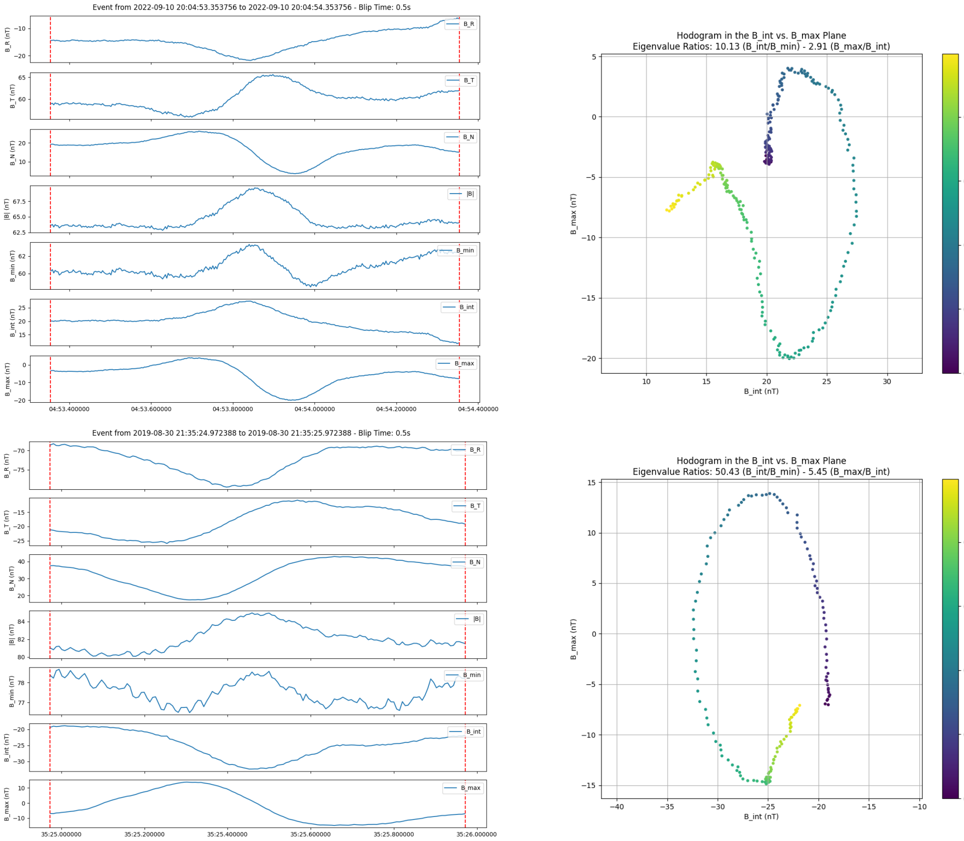


Figure 12: Examples of elliptical-shaped hodograms. The upper panel illustrates an event from the 2022 PSP dataset, while the lower panel depicts an event identified by the RF classifier in the unseen 2019 PSP dataset. Left panels display the RTN magnetic field profiles and the associated MVA profiles over time, while right panels present the corresponding hodograms, plotting B_{int} v.s. B_{max}.

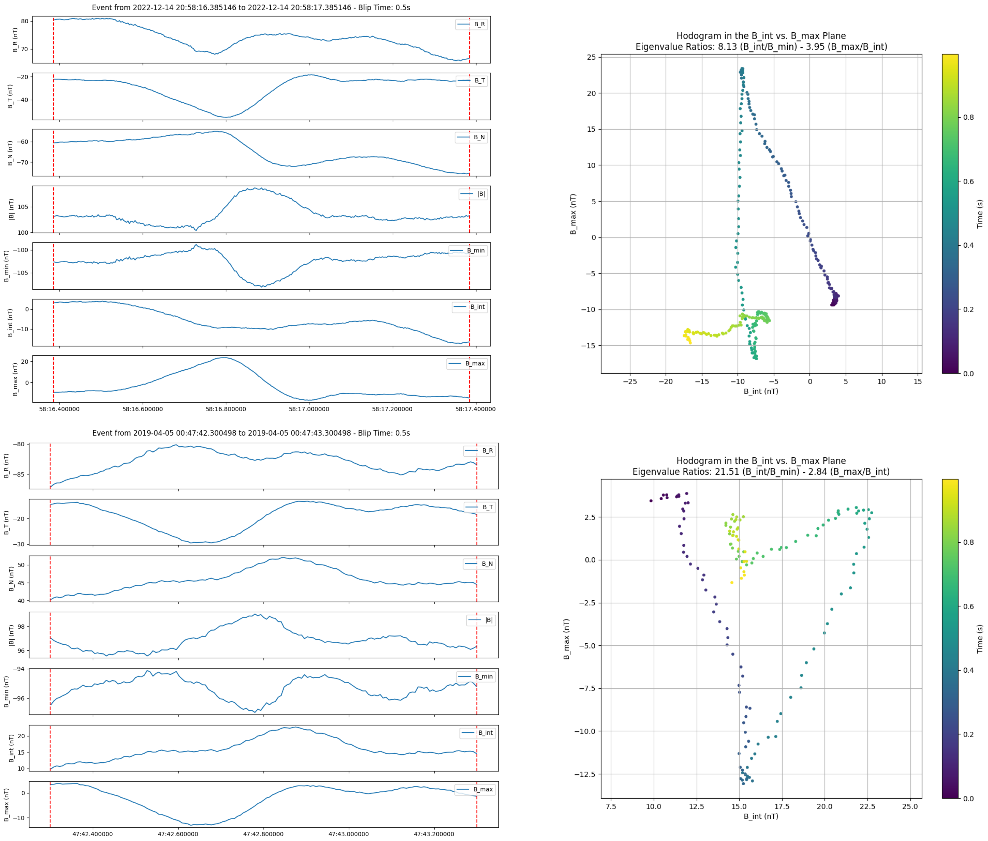


Figure 13: Examples of triangle-shaped hodograms. The upper panel represents an event from the 2022 PSP dataset, while the lower panel shows an event identified by the RF classifier from the unseen 2019 PSP dataset. Left panels display the RTN magnetic field profiles and the associated MVA profiles over time, while right panels present the corresponding hodograms, plotting B_{int} v.s. B_{max}.

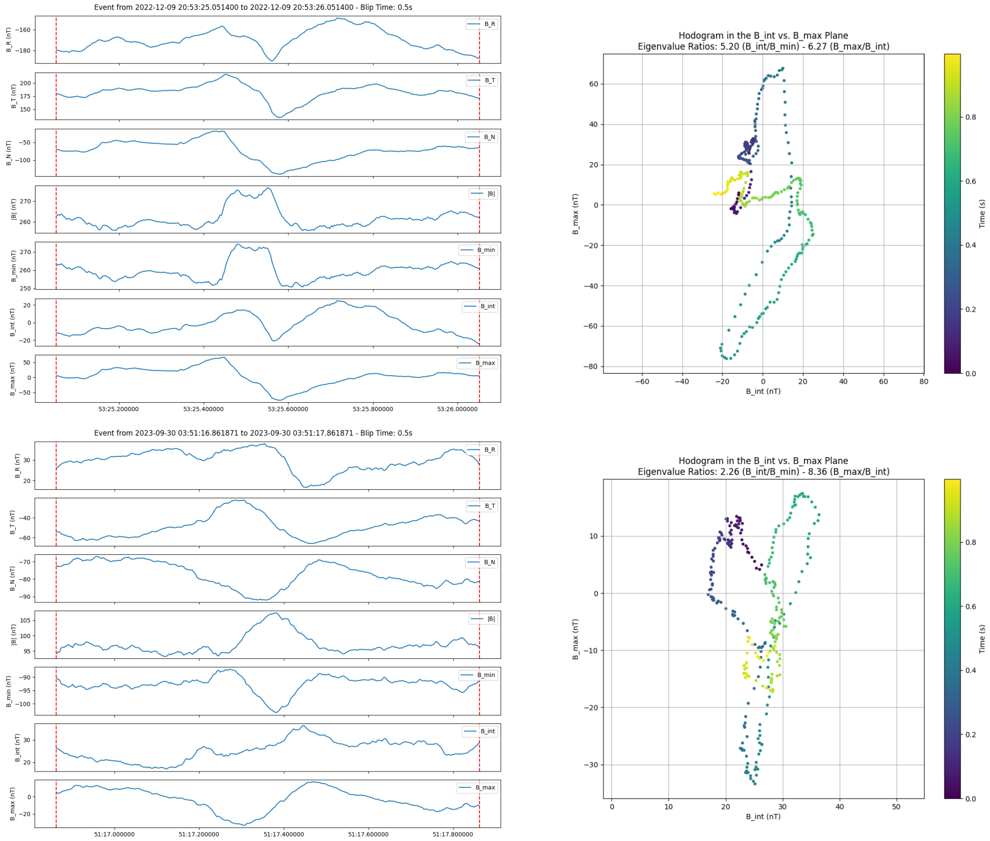


Figure 14: Examples of irregular-shaped hodograms. The upper panel shows an event from the 2022 PSP dataset, while the lower panel shows an event identified by the RF classifier from the unseen 2023 PSP dataset. Left panels display the RTN magnetic field profiles and the associated MVA profiles over time, while right panels present the corresponding hodograms, plotting B_{int} v.s. B_{max}.