The Astrophysical Journal, 961:81 (11pp), 2024 January 20 https://doi.org/10.3847/1538-4357/ad0c52

A Machine Learning Approach to Understanding the Physical Properties of Magnetic Flux Ropes in the Solar Wind at 1 au

Hameedullah Farooki[1,2,3], Yasser Abduallah[1,2], Sung Jun Noh[3,4], Hyomin Kim[1,3], George Bizos[2], Youra Shin[3], Jason T. L. Wang[1,2], and Haimin Wang[1,3,5]
[1] Institute for Space Weather Sciences, New Jersey Institute of Technology, University Heights, Newark, NJ, USA; haf5@njit.edu
[2] Department of Computer Science, New Jersey Institute of Technology, University Heights, Newark, NJ, USA
[3] Center for Solar-Terrestrial Research, New Jersey Institute of Technology, University Heights, Newark, NJ, USA
[4] Now at Los Alamos National Laboratory, Los Alamos, NM, USA
[5] Big Bear Solar Observatory, New Jersey Institute of Technology, 40386 North Shore Lane, Big Bear City, CA 92314, USA
Received 2023 February 23; revised 2023 October 24; accepted 2023 November 12; published 2024 January 16

Abstract

Interplanetary magnetic flux ropes (MFRs) are commonly observed structures in the solar wind, categorized as magnetic clouds (MCs) and small-scale MFRs (SMFRs) depending on whether they are associated with coronal mass ejections. We apply machine learning to systematically compare SMFRs, MCs, and ambient solar wind plasma properties. We construct a data set of 3-minute averaged sequential data points of the solar wind's instantaneous bulk fluid plasma properties using about 20 years of measurements from Wind. We label samples by the presence and type of MFRs containing them using a catalog based on Grad–Shafranov (GS) automated detection for SMFRs and NASA's catalog for MCs (with samples in neither labeled non-MFRs). We apply the random forest machine learning algorithm to find which categories can be more easily distinguished and by what features. MCs were distinguished from non-MFRs with an area under the receiver-operator curve (AUC) of 94% and SMFRs with an AUC of 89%, and had distinctive plasma properties. In contrast, while SMFRs were distinguished from non-MFRs with an AUC of 86%, this appears to rely solely on the $\langle B \rangle > 5$ nT threshold applied by the GS catalog. The results indicate that SMFRs have virtually the same plasma properties as the ambient solar wind, unlike the distinct plasma regimes of MCs. We interpret our findings as additional evidence that most SMFRs at 1 au are generated within the solar wind. We also suggest that they should be considered a salient feature of the solar wind's magnetic structure rather than transient events.

Unified Astronomy Thesaurus concepts: Solar wind (1534); Heliosphere (711); Random Forests (1935)

1. Introduction

Magnetic flux ropes (MFRs) are structures of plasma in space characterized by their spiraling magnetic field lines observed throughout the heliosphere with a wide range of shapes and sizes. They are believed to be the core structure of many phenomena in the atmosphere of the Sun, such as prominences/filaments (Gibson 2018; Liu 2020). As observed in the solar wind near Earth, MFRs are usually classified as either magnetic clouds (MCs) or small-scale magnetic flux ropes (SMFRs; Hu et al. 2018). Both MCs and SMFRs have been detected in all locations of the heliosphere at which the necessary measurements are available, from near the Sun (Chen et al. 2020; Zhao et al. 2020), to near the Earth (Moldwin et al. 2000; Hu et al. 2018; Nieves-Chinchilla et al. 2018), and to above the ecliptic plane (Chen et al. 2019) all the way out to 8 au (Chen & Hu 2020). Although they have similar magnetic structures (hence they are both referred to as MFRs) and are both observed in the solar wind near Earth, MCs and SMFRs are very different and may have different origins. MCs are larger and are the internal magnetic structure of interplanetary coronal mass ejections (ICMEs; coronal mass ejections traveling through the solar wind), whereas SMFRs are smaller, found scattered throughout the bulk of the solar wind, and have an uncertain origin (or multiple origins). MCs and SMFRs differ significantly in their plasma properties, e.g., MCs have very low temperature and plasma beta, whereas SMFRs do not always have these properties.

MCs were originally reported with a strict definition by Burlaga et al. (1981), although the usage of the term has become more general over time (Nieves-Chinchilla et al. 2018). Typically observed a few times a month with durations on the order of tens of hours at 1 au, they are widely accepted to approximately be the internal magnetic structure of many, if not all, interplanetary coronal mass ejections (ICMEs; Hu et al. 2014; Chen 2017; Nieves-Chinchilla et al. 2018), which are coronal mass ejections departed from the Sun propagating through interplanetary space (Howard & Tappin 2009; Kilpua et al. 2017). ICMEs have a significant impact on space weather (Baker et al. 2004), and thus there is much interest in their study.

In contrast, SMFRs occur frequently near Earth, with a few hundred observed monthly on average and a strong solar cycle dependency (Hu et al. 2018). Unlike MCs, the origin of SMFRs is not firmly established (Borovsky 2008; Rouillard et al. 2010, 2011; Sanchez-Diaz et al. 2017a, 2017b, 2019; Hu et al. 2018; Chen & Hu 2022). They are also much smaller, with a typical duration of under an hour at 1 au (Hu et al. 2019). Moreover, their scale sizes range multiple orders of magnitude and their physical properties vary with size (Hu et al. 2018). Furthermore, detecting SMFRs is challenging because the observational signatures are not as clear as those of ICMEs. Understanding SMFRs is important due to their ubiquitous presence in the solar wind, potential significance for

larger-scale phenomena such as solar wind acceleration, and their close relationship with MCs. There has been increased interest in detecting and analyzing SMFRs in recent years, particularly near the Sun using new data from the Parker Solar Probe (Chen & Hu 2022).

SMFRs were first presented by Moldwin et al. (1995, 2000), who identified an SMFR from measurements of the Ulysses spacecraft followed by six more SMFRs from Wind and IMP 8. They found that SMFRs are very similar to MCs in magnetic structure but much smaller and with some different plasma properties. Multiple statistical studies were conducted in later years, but they were typically limited to no more than a few hundred samples (Hu et al. 2018). More recently, Zheng et al. (2017) developed a novel automated detection algorithm based on the Grad–Shafranov (GS) reconstruction technique that could detect tens of thousands of SMFRs from about 20 years of Wind observations. This event list was then formally published by Hu et al. (2018). The GS-based automated detection methodology has since been extended to various spacecraft throughout the heliosphere, such as Ulysses and the Parker Solar Probe (Chen et al. 2019, 2020; Chen & Hu 2022).

Machine learning is a part of artificial intelligence with algorithms for fitting models to data that allow computers to learn patterns in the data and produce predictions for new data or provide insights about the current data (Alpaydin 2020). Machine learning has seen increased use in a wide range of disciplines, including space physics (e.g., Liu et al. 2020; Roberts et al. 2020; Raheem et al. 2021; Richardson & Cohen 2021; Vechm & Malaspina 2021; Zewdie 2021; Abduallah et al. 2022), because machine learning is a versatile tool useful when dealing with large amounts of data. However, there have been limited attempts at applying machine learning to studying MFRs in the solar wind. Recently, there have been a few machine learning applications to MCs. For example, dos Santos et al. (2020) used a deep convolutional neural network to detect signatures of simple MCs in ICMEs and Narock et al. (2022) also used a deep convolutional network to identify the axis of a known MC. Nguyen et al. (2019) developed a deep learning method to detect ICMEs in general. Reiss et al. (2021) used machine learning algorithms to predict the B_z component of an ICME. However, there is still much room for further application of machine learning to MCs. Furthermore, to our knowledge, there have been no attempts to apply machine learning to SMFRs.

In this paper, we apply a machine learning technique to better understand the importance of various physical properties during the presence (or lack thereof) of MCs and SMFRs. We do this by using a machine learning algorithm (known as random forests) to learn the probability distribution of a sample corresponding to various MFR categorizations (e.g., SMFR, MC, and non-MFR) based only on point-in-time physical properties. We use the detected events in the existing catalogs as training data. The purpose of our application of machine learning is not to detect MFRs because point-in-time data is insufficient to properly detect MFRs (e.g., detection algorithms use multiple points in time to detect a rotation in the magnetic field direction, not just looking at the direction at a single point in time). Furthermore, both MC and SMFR catalogs and detection algorithms are already available (see Section 2). Instead, our machine learning approach makes it possible to systematically compare the physical properties of the solar wind under various MFR conditions.

The contribution of our paper is different from previous studies. Hu et al. (2018) compared the physical properties of SMFRs in the fast and slow solar wind, and analyzed the overall statistics of SMFR properties. However, a detailed and systematic comparison of SMFR properties to the background solar wind (when no SMFR is detected), as well as between SMFRs and MCs, is still needed and addressed in this paper. Although we use a machine learning algorithm to distinguish between MFRs and non-MFRs, our ultimate purpose is not to detect MFRs but to analyze the differences between MFR categorizations. Previous machine learning-based studies (e.g., Camporeale et al. 2017 and Li et al. 2020) have introduced various methods to accurately classify solar wind regimes based on in situ measurements. However, our study differs in its focus on MFR categorizations (especially SMFR) and its focus on the strength of the physical properties of the solar wind as probability estimators of the categorizations rather than the overall accuracy of the classifier.

The rest of this paper is structured as follows. In Section 2, we describe in detail the input data and MFR event lists that we utilize. We also provide a brief statistical analysis of the data set. We then describe the machine learning techniques that we use in Section 3, including the classification algorithm and feature selection procedure. Next, we present the results of our experiment in Section 4. Finally, in Section 5, we discuss the results and prospectives for future work.

2. Data

2.1. Input Features

The fundamental physical aspects of the solar wind that we analyze are the magnetic field and plasma properties. The magnetic field is the most relevant due to its fundamental role in defining what an MFR is (i.e., a plasma structure with twisted magnetic field lines), although the single timestamp measurements that we use do not individually contain information about the spatial structure or temporal change. The atomic composition of the solar wind plasma is mostly of protons and then of alpha particles, so we narrow down our focus to those two. Since the data availability and count statistics for protons (which dominate the solar wind) are higher than those of alpha particles, we only use proton parameters rather than alpha particle parameters, other than the alpha/proton number ratio, to indicate the abundance of alpha particles. Because of data gaps in the electron measurements, we opted to leave the electron data out of this study.

All input features come from observations from the Wind spacecraft provided by NASA's Space Physics Data Facility. The time period of the data is from 1996 to 2016. The magnetic field parameters are retrieved from Wind's magnetic field instrument (MFI) (Lepping et al. 1995) at 1 minute cadence, and the plasma parameters are retrieved from the Solar Wind Experiment (SWE) (Ogilvie et al. 1995). We use vector parameters in the Geocentric Solar Ecliptic (GSE) coordinate system. The magnetic field parameters are the three components of the magnetic field vector $\boldsymbol{B}$ (B_x, B_y, and B_z) converted to altitude and azimuth (B_θ and B_ϕ), as well as the magnetic field strength B, which is provided by the MFI data averaged from higher cadence measurements. The plasma parameters are the three components of the proton bulk flow velocity vector $\boldsymbol{u}_\mathrm{p}$ ($u_{\mathrm{p},x}$, $u_{\mathrm{p},y}$, and $u_{\mathrm{p},z}$) converted to altitude and azimuth ($u_{\mathrm{p},\theta}$ and $u_{\mathrm{p},\phi}$), proton bulk speed u_p, proton number density n_p, and the

alpha/proton ratio n_α/n_p. Rather than including the absolute value of the proton temperature, we include the ratio between the observed and expected temperatures based on the proton speed (Richardson & Cane 1995). Although we ignore temporal variations, we included the directional components of the vector quantities because there may be an overall tendency to different directions. We converted to spherical coordinates to separate the magnitude from the direction. Since the SWE dataset has an inconsistent cadence, all of the data is resampled to 3-minute averages. We excluded all data points in which any of the features had no data in the resampling bins.

In addition to the fundamental parameters, relevant derived parameters are added to the data set in case they provide a stronger relationship. Proton gas pressure is added because it is an important property of the solar wind plasma:

$$P_{\mathrm{gas,p}} = n_p k_B T_p \tag{1}$$

where k_B is the Boltzmann constant. Another form of plasma pressure is the dynamic pressure from the motion of the protons:

$$P_{\mathrm{dyn,p}} = \frac{1}{2} m_p n_p u_p^2 \tag{2}$$

where m_p is proton mass. The proton beta is a significant parameter because one of the most important properties of MFRs, especially larger ones, is that they tend to have a low beta (Klein & Burlaga 1982; Hu et al. 2018). The proton beta is calculated using the gas pressure and magnetic field strength:

$$\beta_p = \frac{P_{\mathrm{gas,p}}}{B^2/(2\mu_0)} \tag{3}$$

Finally, two other plasma properties important in the solar wind are Alfvén velocity v_A and Alfvén Mach number $M_A = u_p/v_A$. v_A plays an important role in magnetohydrodynamic fluctuations (Mullan & Smith 2006). It is conceivable that differences in the density and magnetic field of MFRs may result in significant differences in v_A. This may also affect M_A, which is an important indicator of the level of dominance of the magnetic field, plays an important role in shocks (Sundberg et al. 2017), and has an impact on the magnetosphere (Lavraud et al. 2013). We include them both, calculating the Alfvén velocity with the equation:

$$v_A = \frac{B}{\sqrt{2 m_p n_p \mu_0}} \tag{4}$$

2.2. Output Labels

The labels are based on separate sources for MCs and SMFRs. MC labels come from the ICME catalog developed by Nieves-Chinchilla et al. (2018), which is available at https://wind.nasa.gov/ICME_catalog/ICME_catalog_viewer.php. This catalog is based on previous catalogs, as well as additional visual inspection. Most of the ICMEs have either well-defined MCs or similar structures (Nieves-Chinchilla et al. 2018), so we simply use the magnetic object start and end times provided by the catalog, regardless of the magnetic object type.

SMFRs come from the event list on https://fluxrope.info developed by Hu et al. (2018) using the GS automated detection algorithm. We use this source because it is the most comprehensive list of SMFRs detected by Wind. However, we found that most MCs were observed as a series of shorter SMFRs due to the limited window sizes employed by Hu et al. (2018). Therefore, we exclude events overlapping with MCs (approximately 5% of all SMFR events).

The SMFRs were identified by Hu et al. (2018) by trying to identify flux rope structures using segments of the time series observations of the solar wind from Wind. In contrast, this paper focuses on the properties of individual data points, not structures measured over time, to focus on the physical properties associated with MFRs. However, there is a caveat—to avoid detecting mere fluctuations as SMFRs, Hu et al. (2018) excluded SMFR candidates with an average magnetic field strength below 5 nT. As will become apparent later, this poses an issue for analyzing the physical properties of SMFRs.

We distinguish between SMFRs of durations under one hour (labeled short SMFR or SSMFR) and over one hour (labeled long SMFR or LSMFR). The reason for this distinction is that most SMFRs are under an hour in duration, so long SMFRs are unusual. It is possible that LSMFRs and SSMFRs have different origins and impacts on the geospace system. Furthermore, Hu et al. (2018) have noted physical differences between short SMFRs and long SMFRs. Finally, long SMFRs may be expected to have more in common with MCs than short SMFRs due to their more comparable size. Therefore, it is interesting to compare our results for short SMFRs and long SMFRs.

To generate the binary labels from the combined event lists, we mark each timestamp contained within any MFR from either list as positive and the rest as negative. We also add categorical labels describing whether the MFR of a positive timestamp was an SMFR or an MC and a duration label providing the MFR's duration. These labels are not directly used by the machine learning models, but rather they are used to generate the various binary classification tasks described in Section 3 before being removed from the data set.

All of the samples that were not included in either catalog were marked as non-MFRs. Therefore, a limitation of this study is that we assume that the event lists that we use are comprehensive and exclusive. In other words, we assume that all of the events really were MFRs with correct and precise time boundaries, and that all of the time periods with no events in either list contained no MFR whatsoever. Nevertheless, the overall differences in statistical properties should be reasonably reliable assuming that the event lists are mostly correct and complete. As we will see, the main impact is that the difference between non-MFRs and SMFRs is difficult to determine.

Figure 1 (left-hand panels) shows SMFRs on the date 2016 January 18. The shaded regions contain SMFRs according to the catalog. Each feature is plotted in its own subplot. Our data set contains not entire regions but individual data points, which are evenly spaced. For a given sample, there are 15 values corresponding to the 15 features and the aforementioned labels based on what region the sample's timestamp falls under. Figure 1 (right-hand panels) shows an MC that was observed between dates 2016 January 19 and 2016 January 20 in the blue-shaded region. Before and after the MC are SMFRs. In between the MC and SMFRs are unshaded regions, which were not included in either catalog. The samples falling under these regions are assumed to contain no MFR and are thus marked non-MFRs.

2.3. Statistical Analysis

The final data set contains 15 features (described previously) and 2.8×10^6 samples. Of these, 1.9×10^6 (69%) contain

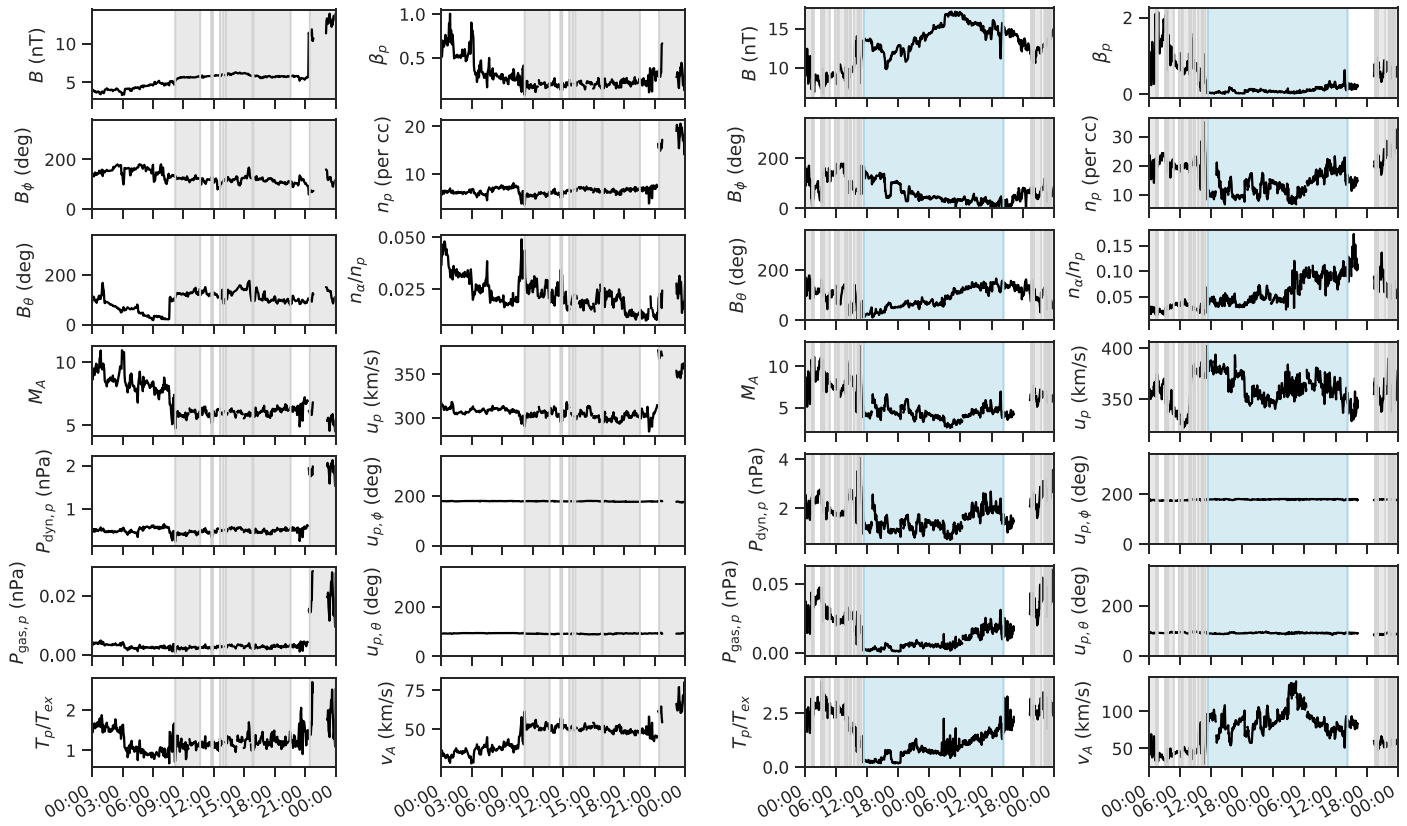

Figure 1. (Left-hand panels) Data for the date 2016 January 18. Shaded regions are SMFR intervals. Data gaps in individual values are represented by breaks in the line plots. (Right-hand panels) Data for dates 2016 January 19 and 2016 January 20. Same format as the left-hand panels, except that the region shaded blue contains an MC.

no MFR, 7.7×10^5 (31%) contain any MFR, 7.3×10^5 (26% of the full data set) contain an SMFR (including 3.0×10^5 SSMFR and 4.3×10^5 LSMFR), and 1.4×10^5 (5% of the full data set) contain an MC. The total number of MFRs is 63807. In total, 63461 of them are SMFRs (from the GS-based list, not overlapping with an MC) and the remaining 346 are MCs (from the ICME list). Of the SMFRs, 51,574 are SSMFRs and 11,887 are LSMFRs.

Considering correlations within the data set is important, both for understanding the data and because correlations can impact feature importance scores (since a machine learning model could use a threshold on either of two variables that have a strong correlation to get the same split of data samples). Figure 2 shows a heatmap of the pairwise correlation matrix of all the features. The correlations are calculated using the Spearman correlation metric, which is similar to the Pearson correlation metric but shows any monotonically increasing (positive correlation)/ decreasing (negative correlation) relationship rather than only linear relationships. The correlation matrix is symmetric, and so it is the same across the diagonal. The diagonal contains all ones because each feature is perfectly correlated with itself. While many feature pairs have almost no correlation, there are some significant correlations. The strongest ones should be kept in mind when interpreting the results.

3. Methodology

3.1. Classification Tasks

We define six binary classification tasks to be solved by our model based on the data: SSMFR-NMFR, LSMFR-NMFR, SSMFR-LSMFR, MC-NMFR, MC-SSMFR, and MC-LSMFR. Each task consists of estimating the likelihood of a sample's true label being positive based on a subset of the data marked as positive merged with another subset of the data marked as negative. For example, SSMFR-NMFR has samples in SMFRs of duration under 1 hr as the positive class and non-MFR samples as the negative class. Unmarked labels are excluded from the data set for the corresponding task. The different tasks extensively explore the MFR categorization of a point in time and are defined so that every pair is compared.

3.2. Algorithm

We adopt the random forest classification algorithm (Breiman et al. 1984) to generate our machine learning model. This algorithm is well-suited because it tries random subsets of the features on random subsets of the data to build an ensemble model. The random sampling of features is used to thoroughly explore different possible feature combinations to give them a chance to demonstrate their predictive power and importance with respect to the labels. Meanwhile, the random subsetting of the data, or bootstrapping, is used to ensure that the trees are more diverse. It is widely used in many domains and has been used successfully for space weather science in particular. For example, Liu et al. (2017) used a random forest in order to forecast solar flares in solar active regions.

In simple terms, the random forest algorithm learns by using a training set of input data vectors and corresponding output categorizations, or labels, to generate an ensemble of randomized decision trees that collectively vote on the correct label of a new input. The model is trained as follows: Given a training set

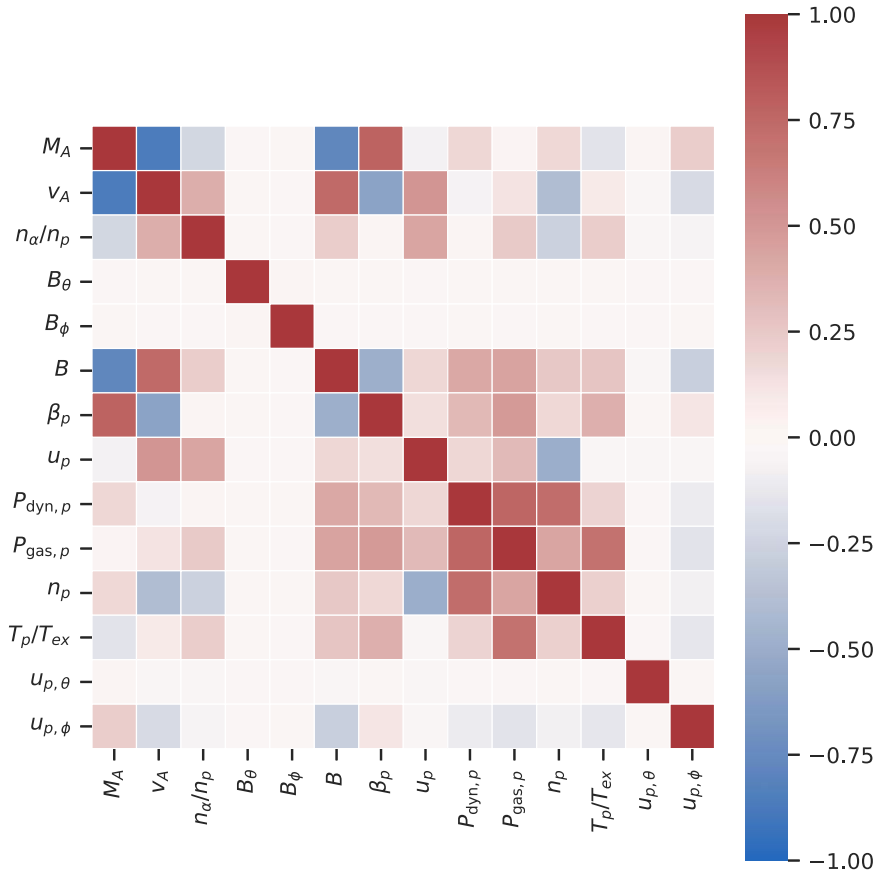

Figure 2. Correlation heatmap, which shows Spearman correlation between features including raw measurements and derived parameters. Higher positive values indicate a co-increasing relationship, whereas lower negative values indicate an inverse relationship. Near-zero indicates a lack of any monotonously decreasing or increasing relationship.

consisting of N vectors with M features $X = x_1, x_2, \cdots, x_{N-1}, x_N$ with labels $Y = y_1, y_2, \cdots, y_{N-1}, y_N$, generate B binary trees T_1, T_2, $\cdots$, T_{B-1}, T_B according to the following algorithm:

1. Sample N elements (with replacement) of the training set, denoted (X_i, Y_i).
2. Train binary decision tree T_i on (X_i, Y_i). At each split, a subset of $\sqrt{M}$ features are selected to compare. The optimal feature for splitting is selected by maximizing the resulting information gain or minimizing the Gini impurity measure.

In our case, we limit the number of trees to 64, the tree depth to eight, and the maximum number of samples used for training each tree to 10%. Tree splitting is based on the Gini impurity. When training the model, we undersample the negative class, i.e., we select a random subset of the negative samples to match the number of positive samples so that the model is not biased toward negative samples.

3.3. Hyperparameter Tuning

Several parameters must be set before training a random forest model, otherwise known as hyperparameters. Although random forests are generally not overly sensitive to the chosen hyperparameters, the most important are the number of trees and number of features used per tree. Furthermore, although not quite a hyperparameter, a tradeoff between the model complexity and preventing overfitting must be made. This is done by limiting the size of the trees, such as by limiting the tree depth. Lower tree depth also entails faster training.

To optimize the hyperparameters, we temporarily set aside 10% of the training sets as validation sets. We evaluated how the varying parameters affected all of the tasks. We found that the default recommendation for the number of features per tree was optimal. Additionally, we experimented with numbers of trees including 32, 64, and 128. We found that 64 trees provided optimal performance without unnecessary computational overhead. Furthermore, we found that a maximum tree depth of eight provided optimal performance without overfitting. Finally, we found that only 10% of the samples (randomly selected with replacement) per tree were needed due to the large size of our data set. Therefore, we applied this limit to decrease the training time. An added benefit is that there is more disparity between the samples used to train each tree of the random forests, thus increasing their diversity.

3.4. Feature Importances

After building a model using the random forest algorithm, we use Gini importance to rank the importance of each feature for that model (Breiman et al. 1984). To calculate the importance of each feature, we start by calculating the non-normalized importance $\hat{I}_{f,i}$ of each feature f for each individual tree i. We calculate the sum of the gain G_j (calculated by the random forest algorithm) weighted by the number of samples n_j reaching each node j out of the tree's nodes that split on

feature f:

$$\hat{I}_{f,i} = \sum_j n_j G_j$$

We then normalize the feature importances by dividing them by the sum of all of the importances:

$$I_{f,i} = \frac{\hat{I}_{f,i}}{\sum_i \hat{I}_{f,i}}$$

Finally, we take the mean of the normalized importances from each tree to get the overall forest's importance levels for each feature:

$$I_f = \frac{1}{B}\sum_i \hat{I}_{f,i}$$

3.5. Feature Selection

To evaluate the relevance of the features to each task, we evaluate the model's performance for each task on a subset of the data set that was set aside for testing, which is not used for training or hyperparameter tuning. An important consideration is how to split the data set into testing and training sets. If we use random sampling, then that results in the test set being drawn from the same time period as the training set, which can potentially allow the model to overfit to the time periods without reducing the performance score on the test set. Meanwhile, the solar wind properties and flux rope occurrence rate change throughout the solar cycle, so the test set should include samples from throughout the data set. Therefore, we split the data into 20 sequential segments and take the first 20% of each of those 20 segments for the test set, and use the remainder to build the training set, which contains 80% of the full data set.

Using the Gini metric of feature importance, we perform feature selection by iteratively eliminating features in order of least to most important, i.e., for each task, we remove the least important feature from the data set, and we then re-train and re-evaluate the model with the remaining features. The performance is evaluated using the area under the receiver-operator curve (AUC) score, which is equivalent to the probability that a positive sample will receive a higher model output than a negative sample (Hanley & McNeil 1982). We continue to do this (without recalculating the feature importances) until only the most important feature remains. We then plot this result to show the AUC score for each number of features, demonstrating how the model's performance decreases when increasingly important features are removed.

A well-known limitation of using Gini importance for feature ranking is that if some of the features are correlated, then all of the importance can be given to one of the features at the expense of the other. A commonly used alternative is permutation importance, which measures the decrease in the model's performance if a given feature's values are shuffled randomly. We have opted to use Gini importance for the following reasons. First, permutation importance is only meaningful if the model has high performance, but for some of the cases in this paper the random forest classifier was unable to distinguish between the two categories with good performance. In contrast, Gini importance is still meaningful for a poorly performing model because it measures the relative information gained from a particular feature. Second, permutation performance can sometimes make the issue of correlations even worse by reducing the importance of both correlated features. Finally, by experimenting with the use of permutation importance instead of Gini importance, we have found that for the data used in this study the results do not differ significantly and the conclusions are unaffected. While more sophisticated methods exist to determine feature importances for correlated features, the number of features in our study is not that large and the correlations between them can be understood physically and kept in mind when interpreting the results. Therefore, we have opted not to include the results with permutation importance in the paper. Instead, we discuss the effect of the correlations by physical reasoning and use them to reach conclusions about the most significant features in Section 4.

One issue that may arise when interpreting the feature selection results is determining how many features one needs to avoid losing a statistically significant amount of performance. It is difficult to evaluate this visually when the slope is not steep. Our solution to this is as follows: We split the training set into 10 disjoint training folds of equal size using the same time segmenting procedure as before. Furthermore, we split the test set into 10 disjoint testing folds of equal size. We then iterate 10 times. For each iteration i, we train the model on the ith training fold. Then, once the model is trained, we perform a nested iteration 10 times. For each nested iteration j, we test the model on the jth testing fold. We calculate the AUC score for iteration ij and store it. At the end, there are $10^2 = 100$ AUC scores. This is repeated with the top A features and top B features, yielding two sets of 100 AUC scores for statistical comparison. We begin with $A = 1$ and $B = 2$, then $A = 2$ and $B = 3$, and so on. Each time, we perform a Wilcoxon signed-rank test (Wilcoxon 1947) to determine whether there is a statistically significant improvement in performance between using the top A features and the top B features. We continue doing this until we encounter a p-value of greater than 5%. For example, if this occurs when comparing the top five versus the top six features for a particular task, then the top five features are considered significant for that task.

4. Results

We trained the classification model for each of the six tasks using the random forest algorithm and their respective training sets. The models themselves consist of many fairly deep trees, so it would be very difficult to visualize them directly. In Figure 3, we show an example of an individual decision tree from task SSMFR-NMFR. This illustrates how the decision tree process works for individual trees. The overall random forest is, however, based on an ensemble of many trees. Hence, we use statistical methods below to understand the results.

We present a detailed breakdown of the performance of each task across the different MFR classes in the rest of Table 1. The table is generated as follows. We first iterate through each class (NMFR, SSMFR, LSMFR, SMFR, MC, MFR). Then, using the initial models trained on the training sets with all of the features, we predict the class of each sample in the test set belonging to the current class. We calculate the average predicted class $\langle y' \rangle = \frac{1}{N}\sum y_i'$ where y_i' is 1 if the model classifies a sample as positive and 0 if it classifies a sample as negative. For tasks where the current class is a subset of the positive class, we use $\langle y' \rangle$ as the accuracy. If the current class is a subset of the negative class, then we use $1 - \langle y' \rangle$ as the

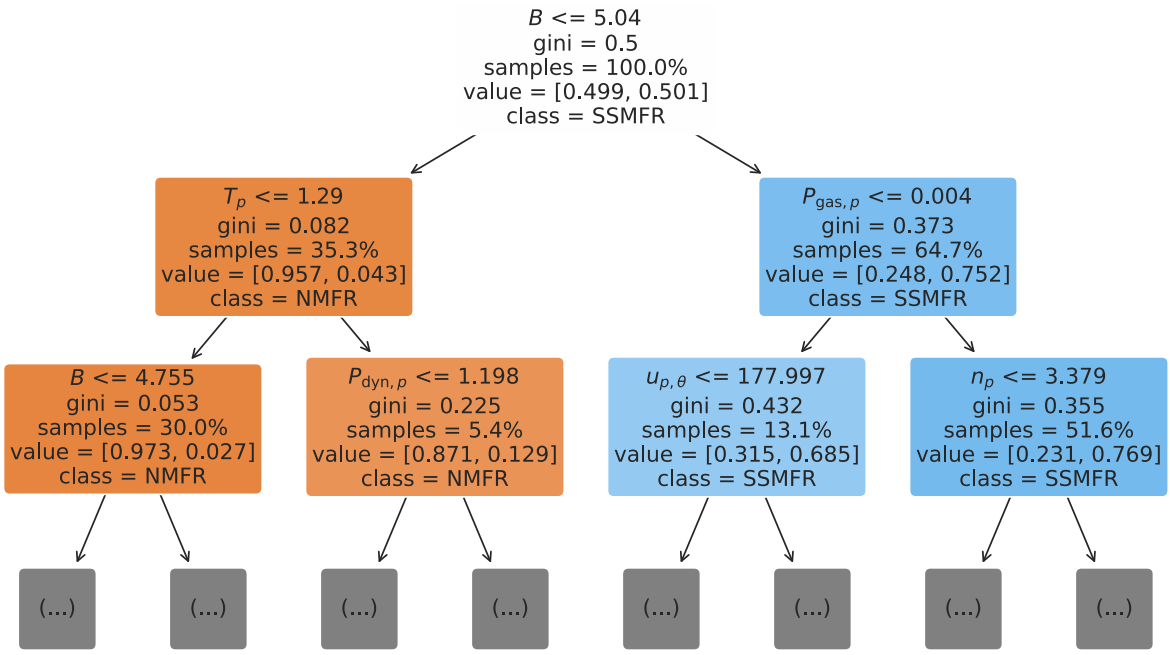

Figure 3. Examples of a decision tree from SSMFR-NMFR. Only the first few levels of the tree are shown. Each task's random forest has many trees, which was selected because it started with the top feature for the respective test. The gini metric shows the impurity after the previous split. Branches to the left-hand result when the condition is satisfied. The "value" shows the percentage of NMFRs (left-hand) and percentage of SMFRs (right-hand) assuming that the samples start off balanced.

Table 1
Cross-task Performance

Class (%)	SSMFR-NMFR (%)	LSMFR-NMFR (%)	SSMFR-LSMFR (%)	MC-NMFR (%)	MC-SSMFR (%)	MC-LSMFR (%)
NMFR	66.8	66.5	⋯	88.2	⋯	⋯
SSMFR	97.0	⋯	64.0	⋯	87.6	⋯
LSMFR	⋯	93.1	63.9	⋯	⋯	79.0
MC	⋯	⋯	⋯	83.5	82.8	78.5
AUC Score	86.1	85.2	66.2	94.1	92.7	86.4

accuracy. If the current class is not a subset of either one, then there is no meaningful accuracy, so we put dashed lines. The results are recorded in Table 1. Additionally, the resulting AUC scores on the test sets are displayed in the last row of Table 1 With the exception of SSMFR-LSMFR, all of them had around 90% AUC, with MC-NMFR exceeding 94%. The MC-related tasks performed better than the SMFR-related tasks. Higher AUC scores indicate that the model is able to assign a higher prediction value to a true sample over a negative sample more often, and thus the features given to it have more predictive power. The baseline AUC score for random guessing is 0.5, so in all cases, the model had significant (if not good) performance. The fact that the model can give meaningful probability estimates indicates that there is a significant correspondence between the features that it found and the labels of each task.

The poor performance of SSMFR-LSMFR is important. If big SMFRs tended to be more similar to MCs, then there should have been a big physical difference between LSMFR and SSMFR, and SSMFR-LSMFR should have scored well. In fact, MC-LSMFR does perform worse than MC-SSMFR, but SSMFR-LSMFR performs worse than either one of them, suggesting that big SMFRs have more in common with small SMFRs than MCs. The physical properties exhibit only slight differences, as we will see below. We will also explore the reasons why MC-NMFR, SSMFR-NMFR, and LSMFR-NMFR perform well.

After training each task's model, we calculate the feature importances for each task, and then use it to rank the features, recorded in Table 2. Finally, we generate the feature selection curves for each task. The least important feature is dropped iteratively until only the most important feature remains, and the model is re-trained with each subset of features and the AUC score is calculated. Figure 4(a) shows the plot of the results for each task. To determine which features were significant for each task, we performed the Wilcoxon tests in the manner described in Section 3.5. Using this information, we colored the significant features bold in Table 2. We also sorted the features by their average importance across tasks and then conducted the same process except using this new cross-task ordering. The resulting feature elimination plot is Figure 4(b). The top five features across tasks were B, β_p, T_p/T_{exp}, M_A, and P_{gas}.

The feature ranking tells us which features are significant for distinguishing a sample's class, but not what values correspond to a higher probability. These can be understood with histograms and partial dependency plots (PDPs), plotted in Figure 5. To generate the PDPs, each model is trained with all features using the task's training set, and then 10,000 random samples from the task's test set are selected. Take magnetic field strength B as an example. For each point on the PDP, the

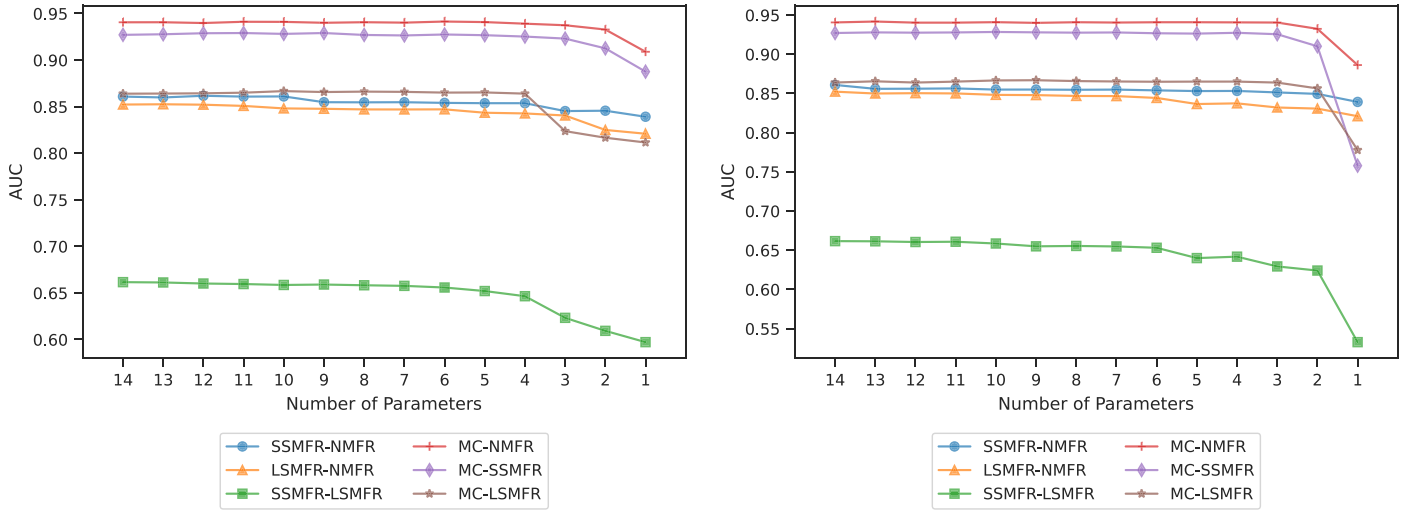

Figure 4. (a) Feature elimination results generated using each individual task's feature ranking. (b) Same as (a), except rather than dropping features based on each task's ranking, we drop in order of average importance across tasks.

magnetic field strength B of each selected sample is changed to a new value B', and then all of the samples are passed through the task's model. The mean resulting probability prediction is represented by a point on the PDP. The change of the value of the PDP shows how the model's output depends on a given parameter.

For both SSMFR-NMFR and LSMFR-NMFR, the top feature is the magnetic field strength B. This implies that the magnetic field strength stands out the most during the SMFRs in the event list compared to other features. The comparison of their distributions in Figure 5 reveals a sharp change centered at 5 nT. Only 2.7% of samples with $B < 5$ nT contain an SMFR, whereas 47.6% of samples with $B \geqslant 5$ nT contain an SMFR. Apparently, this is due to the fact that Hu et al. (2018) excluded any MFR event candidates with an average magnetic field strength below a 5 nT threshold to exclude random fluctuations. We have also included the distribution for the non-MFR samples with the same threshold applied, which results in a very similar distribution to the SMFR distribution. The PDP shows a very sharp increase after 5 nT from approximately 0.1 or less to a nearly constant value for SMFR tasks. In contrast to the PDP for the previously discussed tasks, the MC tasks have a smooth increase of probability with magnetic field strength. Interestingly, MC-SSMFR and MC-LSMFR have an increase in probability below 5 nT. This suggests that at values below the threshold, the model gives a higher chance of the sample being an MC than an SMFR, even though MCs tend to have higher magnetic field strength than SMFRs. Overall, it seems that there may be many SMFRs with lower magnetic field strength and that the threshold has imposed a selection bias.

For the MC-related tasks, the most important feature is proton beta. In fact, judging by the PDP, $\beta_{\rm p}$ is the main factor in distinguishing MCs from other categories. It is well-known that the proton beta tends to be very low for ICMEs (e.g., Liu et al. 2005). MCs tend to have near-0 values, which only occur rarely in the background solar wind. However, the tail of the distribution cannot be ignored because observational studies show that not all MCs have low proton beta for their entire duration (e.g., Pal et al. 2022).

SSMFRs and LSMFRs also tend to have lower proton beta than NMFRs, although the difference is not nearly as pronounced as in the case of MCs. The distribution of SMFR proton beta has been analyzed by Hu et al. (2018). The smaller

Table 2
Feature Importances (Significant Features in Bold)

Rank	SSMFR-NMFR	LSMFR-NMFR	SSMFR-LSMFR	MC-NMFR	MC-SSMFR	MC-LSMFR
1	$\boldsymbol{B}$	$\boldsymbol{B}$	$\boldsymbol{\beta_p}$	$\boldsymbol{\beta_p}$	$\boldsymbol{\beta_p}$	$\boldsymbol{\beta_p}$
2	$\boldsymbol{M_A}$	$\boldsymbol{M_A}$	$\boldsymbol{T_p/T_{ex}}$	$\boldsymbol{B}$	$\boldsymbol{T_p/T_{ex}}$	$\boldsymbol{T_p/T_{ex}}$
3	$v_{\rm A}$	$\boldsymbol{v_A}$	$\boldsymbol{P_{gas,p}}$	$\boldsymbol{M_A}$	$\boldsymbol{P_{gas,p}}$	$\boldsymbol{M_A}$
4	$P_{{\rm gas},p}$	$\boldsymbol{P_{dyn,p}}$	$\boldsymbol{u_p}$	$\boldsymbol{v_A}$	$\boldsymbol{M_A}$	$\boldsymbol{B}$
5	$P_{{\rm dyn},p}$	$\boldsymbol{n_p}$	$\boldsymbol{B}$	$\boldsymbol{T_p/T_{ex}}$	$\boldsymbol{B}$	$\boldsymbol{P_{gas,p}}$
6	$n_{\rm p}$	$\boldsymbol{\beta_p}$	$\boldsymbol{n_\alpha/n_p}$	$P_{{\rm gas},p}$	$\boldsymbol{v_A}$	$v_{\rm A}$
7	$T_{\rm p}/T_{\rm ex}$	$P_{{\rm gas},p}$	$\boldsymbol{v_A}$	$n_{\rm p}$	$P_{{\rm dyn},p}$	$n_{\rm p}$
8	$\beta_{\rm p}$	$T_{\rm p}/T_{\rm ex}$	$\boldsymbol{n_p}$	$P_{{\rm dyn},p}$	$n_{\rm p}$	$P_{{\rm dyn},p}$
9	$u_{\rm p}$	$u_{\rm p}$	$M_{\rm A}$	$u_{\rm p}$	$n_\alpha/n_{\rm p}$	$n_\alpha/n_{\rm p}$
10	B_θ	$u_{{\rm p},\phi}$	$P_{{\rm dyn},p}$	$n_\alpha/n_{\rm p}$	$u_{\rm p}$	$u_{\rm p}$
11	$u_{{\rm p},\theta}$	B_θ	B_ϕ	B_θ	$u_{{\rm p},\phi}$	B_ϕ
12	$u_{{\rm p},\phi}$	$n_\alpha/n_{\rm p}$	$u_{{\rm p},\theta}$	B_ϕ	$u_{{\rm p},\theta}$	$u_{{\rm p},\theta}$
13	$n_\alpha/n_{\rm p}$	B_ϕ	$u_{{\rm p},\phi}$	$u_{{\rm p},\phi}$	B_ϕ	$u_{{\rm p},\phi}$
14	B_ϕ	$u_{{\rm p},\theta}$	B_θ	$u_{{\rm p},\theta}$	B_θ	B_θ

difference between MC and SMFR proton beta compared to MC and non-MFR may explain the lower AUC score for MC-SSMFR and MC-LSMFR. Likewise, it appears that LSMFRs have slightly lower $\beta_{\rm p}$ than SSMFRs, so MC-LSMFR performs worse than MC-SSMFR. However, the shape of the distribution for LSMFRs, SSMFRs, and NMFRs is virtually the same, whereas MC-LSMFR has a completely different distribution shape, so this does not necessarily indicate a similarity between the plasma properties of LSMFRs and MCs. In fact, both SSMFRs and LSMFRs have almost the same $\beta_{\rm p}$ distribution as NMFRs that have $B > 5$ nT, suggesting that they both share the distribution of the background solar wind.

Most features besides B do not appear to be significantly different for SMFRs based on the histograms compared to NMFRs (with the 5 nT threshold applied). $M_{\rm A}$ was given high importance for SMFR-related tasks, but does not appear to affect the PDP much compared to B. Since it is highly correlated with B in Figure 2, this is probably what resulted in its high ranking. In contrast, MCs have significantly different shape of the $M_{\rm A}$ distribution. However, since the overall distribution has too much overlap in terms of the range of values, it was not ranked highly in terms of Gini importance.

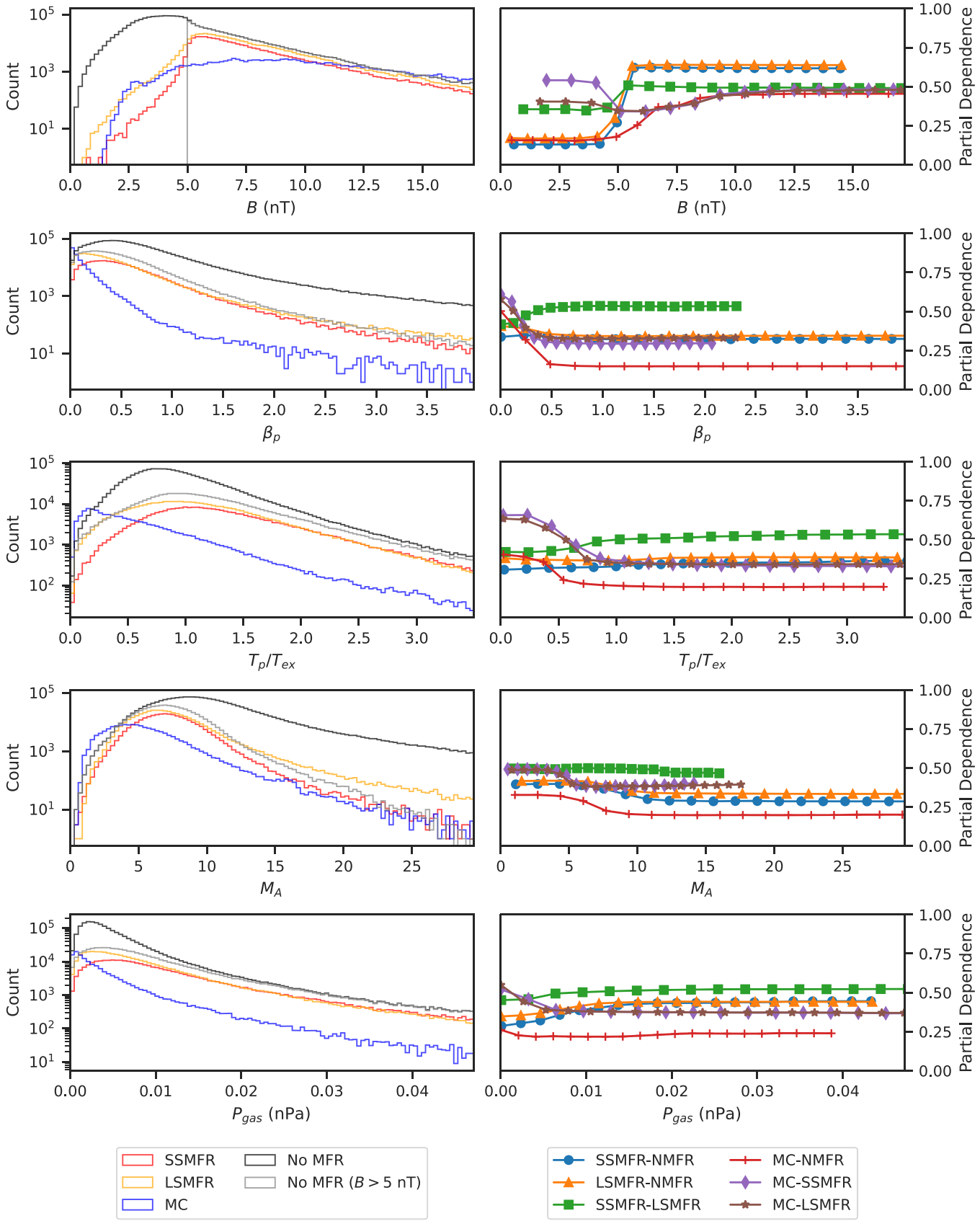

Figure 5. Histogram and partial dependency plots for the top 5 most important features across tasks.

Beta, temperature, and gas pressure appear to play an important role for distinguishing MCs from NMFRs and LSMFRs from SSMFRs. As Hu et al. (2018) pointed out, smaller SMFRs have a wide range of temperatures, whereas larger SMFRs have lower temperatures. This finding is consistent with the PDP, which shows that SSMFRs are more

likely to have higher temperatures than LSMFRs. But even though LSMFRs have low temperature, their temperature distribution is shaped completely differently from MCs. Unlike MCs, the distribution of LSMFR temperature is similar to both SSMFR and NSMFR temperature. Accordingly, temperature only plays a very strong role for MC-related tasks in the PDP. Gas pressure, related to temperature by $P = nk_BT$, tends to be very low for MCs compared to SMFRs, LSMFRs, or NMFRs, which all have a similar distribution of gas pressure. However, larger SMFRs appear to have slightly lower gas pressure.

5. Discussion and Conclusions

In this paper, we ranked the informativeness of various physical properties of the solar wind properties with regards to various MFR-related categorizations (SSMFR, LSMFR, MC, and NMFR) using a machine learning feature selection technique. We first generated a data set using measurements from Wind. We then used six binary classifications to compare the point-in-time physical properties of each unique pair of categories. For each task, we used the random forest algorithm to build an ensemble of decision trees to distinguish the positive from negative samples and then used the normalized average decrease in the impurity of each feature to rank the features. After that, we applied feature selection by plotting the change in model performance with increasingly reduced sets of top features used for training and classification.

Our results demonstrate that MCs tend to have very distinctive solar wind properties compared to the absence of any MFR, such that an excellent probability estimate of a given timestamp corresponding to an MC instead of a non-MFR can be given based on plasma and magnetic field measurements from a single point in time. The proton beta alone is able to give an AUC score of over 0.9 for distinguishing MCs from non-MFRs, whereas the top three features together (β_p, B, and v_A) result in an AUC score of close to 0.94. This makes sense because it is known that intervals of both low β_p and v_A are most common in the solar wind during ICMEs (Gosling & Phan 2013), and that MCs tend to have elevated magnetic field strength B (Nieves-Chinchilla et al. 2018).

We found that SMFRs in the catalog can be distinguished from non-MFRs with significant performance. However, the primary distinguishing feature is just the $B > 5$ nT threshold used by Hu et al. (2018) to make the catalog. As a result, the plasma properties that differ significantly in distribution from the non-MFRs—v_A, M_A, and P_{gas} in particular, are strongly correlated with B. Therefore, it seems that SMFRs share the same distribution as the background solar wind in terms of instantaneous physical properties. A potential area for future research is to detect flux ropes from multiple-spacecraft observations. This would eliminate the need for the threshold because one can be more confident of a detected flux rope not being a mere fluctuation.

Our comparison between SMFRs shorter or longer than an hour in duration—corresponding to small and large: SMFRs tend to be Parker spiral aligned (Hu et al. 2018), so duration is a good proxy for size—suggests that there is almost no different between their physical properties. Even though, as noted by Hu et al. (2018), larger SMFRs have lower proton temperature, proton beta, and gas pressure, the difference is small compared to how low the temperature and beta of MCs is. Larger SMFRs have more in common with smaller SMFRs and the background solar wind than with MCs. The main difference in physical properties between larger SMFRs and smaller SMFRs is that larger SMFRs have slightly lower gas pressure, temperature, and proton beta. However, the distributions of these parameters have the same shape as the distributions of smaller events, whereas both are very different from MCs. Additionally, the magnetic field fluctuates more in smaller flux ropes than larger ones, although that is due to their structure's size and not their plasma properties.

Based on our results, we hypothesize that most SMFRs at 1 au share the same plasma properties as the ambient solar wind. This is in contrast to the typical description of SMFRs as transient phenomena such as MCs, which are distinct plasma regimes in the solar wind. This makes sense considering that the SMFR events in the current catalog with the 5 nT threshold applied already span 26% of the full time, i.e., SMFRs are present at 1 au over one quarter of the time. In the context of the flux-tube picture of the solar wind by Borovsky (2008), this discussion suggests that the flux tubes filling the solar wind are often twisted, which may have important ramifications for the interaction between the solar wind and the magnetosphere. It also provides further support for most SMFRs at 1 au having a local origin within the solar wind, e.g., by turbulence (Hu et al. 2018). However, this study is based on the bulk fluid properties of the solar wind. Analysis of the microscopic properties such as particle distribution functions may yield different conclusions regarding the origin of the SMFRs.

Future works may extend this feature ranking methodology to data from other spacecraft in different parts of the heliosphere and compare the change in results based on variations in ecliptic latitude and distance from the Sun. Furthermore, this methodology can be applied to other phenomena with various in situ time series measurements of which the relative importances are of interest.

Acknowledgments

We thank NASA's Space Physics Data Facility for providing the Wind data, and we thank the teams at the University of Alabama in Huntsville (UAH) and NASA for the SMFR (available at https://fluxrope.info) and MC (available at https://wind.nasa.gov/ICME_catalog/ICME_catalog_viewer.php) catalogs, respectively. We acknowledge the support of NASA grant 80NSSC20K1282 and NSF grants AGS-1927578, AGS-2229064, and OPP-2032421. The figures and results in this study were generated using Python and Python libraries including matplotlib (Hunter 2007), SciKit-Learn (Pedregosa et al. 2011), numpy (Harris et al. 2020), scipy (Virtanen et al. 2020), and pandas (McKinney 2010).

ORCID iDs

Hameedullah Farooki https://orcid.org/0000-0001-7952-8032
Yasser Abduallah https://orcid.org/0000-0003-0792-2270
Sung Jun Noh https://orcid.org/0000-0002-8032-7833
Hyomin Kim https://orcid.org/0000-0002-6350-405X
Youra Shin https://orcid.org/0000-0002-0815-9855
Jason T. L. Wang https://orcid.org/0000-0002-2486-1097
Haimin Wang https://orcid.org/0000-0002-5233-565X

References

Abduallah, Y., Jordanova, V. K., Liu, H., et al. 2022, ApJS, 260, 16
Alpaydin, E. 2020, Introduction to Machine Learning (Cambridge, MA: MIT press)

Baker, D. N., Daly, E., Daglis, I., Kappenman, J. G., & Panasyuk, M. 2004, SpWea, 2, S02004
Borovsky, J. E. 2008, JGRA, 113, A08110
Breiman, L., Friedman, J. H., & Olshen, R. A. 1984, Classification and Regression Trees (Boca Raton, FL: Chapman and Hall/CRC)
Burlaga, L., Sittler, E., Mariani, F., & Schwenn, A. R. 1981, JGR, 86, 6673
Camporeale, E., Carè, A., & Borovsky, J. E. 2017, JGRA, 122, 10910
Chen, J. 2017, PhPl, 24, 090501
Chen, Y., & Hu, Q. 2020, ApJ, 894, 25
Chen, Y., & Hu, Q. 2022, ApJ, 924, 43
Chen, Y., Hu, Q., & le Roux, J. A. 2019, ApJ, 881, 58
Chen, Y., Hu, Q., Zhao, L., et al. 2020, ApJ, 903, 76
dos Santos, L. F., Narock, A., Nieves-Chinchilla, T., Nuñez, M., & Kirk, M. 2020, SoPh, 295, 131
Gibson, S. E. 2018, LRSP, 15, 1
Gosling, J. T., & Phan, T. D. 2013, ApJL, 763, L39
Hanley, J. A., & McNeil, B. J. 1982, Radiology, 143, 29
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Howard, T. A., & Tappin, S. J. 2009, SSRv, 147, 31
Hu, Q., Chen, Y., & le Roux, J. 2019, JPhCS, 1332, 012005
Hu, Q., Qiu, J., Dasgupta, B., Khare, A., & Webb, G. M. 2014, ApJ, 793, 53
Hu, Q., Zheng, J., Chen, Y., le Roux, J., & Zhao, L. 2018, ApJS, 239, 12
Hunter, J. D. 2007, CSE, 9, 90
Kilpua, E., Koskinen, H. E., & Pulkkinen, T. I. 2017, LRSP, 14, 1
Klein, L., & Burlaga, L. 1982, JGR, 87, 613
Lavraud, B., Larroque, E., Budnik, E., et al. 2013, JGRA, 118, 1089
Lepping, R., Acũna, M., Burlaga, L., et al. 1995, SSRv, 71, 207
Li, H., Wang, C., Tu, C., & Xu, F. 2020, E&SS, 7, e2019EA000997
Liu, C., Deng, N., Wang, J. T. L., & Wang, H. 2017, ApJ, 843, 104
Liu, H., Liu, C., Wang, J. T. L., & Wang, H. 2020, ApJ, 890, 12
Liu, R. 2020, RAA, 20, 165
Liu, Y., Richardson, J., & Belcher, J. 2005, P&SS, 53, 3
McKinney, W. 2010, in Proc. 9th Python in Science Conf. (SciPy 2010), ed. S. van der Walt & J. Millman, 56
Moldwin, M., Ford, S., Lepping, R., Slavin, J., & Szabo, A. 2000, GeoRL, 27, 57
Moldwin, M., Phillips, J., Gosling, J., et al. 1995, JGR, 100, 19903
Mullan, D., & Smith, C. 2006, SoPh, 234, 325
Narock, T., Narock, A., Dos Santos, L. F., & Nieves-Chinchilla, T. 2022, FrASS, 9, 838442
Nguyen, G., Aunai, N., Fontaine, D., et al. 2019, ApJ, 874, 145
Nieves-Chinchilla, T., Vourlidas, A., Raymond, J., et al. 2018, SoPh, 293, 25
Ogilvie, K., Chornay, D., Fritzenreiter, R., et al. 1995, SSRv, 71, 55
Pal, S., Lynch, B. J., Good, S. W., et al. 2022, FrASS, 9, 903676
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Raheem, A.-u., Cavus, H., Coban, G. C., et al. 2021, MNRAS, 506, 1916
Reiss, M. A., Möstl, C., Bailey, R. L., et al. 2021, SpWea, 19, e2021SW002859
Richardson, D. K., & Cohen, M. B. 2021, JGRA, 126, e29689
Richardson, I., & Cane, H. 1995, JGR, 100, 23397
Roberts, D. A., Karimabadi, H., Sipes, T., Ko, Y.-K., & Lepri, S. 2020, ApJ, 889, 153
Rouillard, A., Davies, J., Lavraud, B., et al. 2010, JGRA, 115, A04103
Rouillard, A., Sheeley, N., Cooper, T., et al. 2011, ApJ, 734, 7
Sanchez-Diaz, E., Rouillard, A., Lavraud, B., Kilpua, E., & Davies, J. 2019, ApJ, 882, 51
Sanchez-Diaz, E., Rouillard, A. P., Davies, J. A., et al. 2017a, ApJ, 851, 32
Sanchez-Diaz, E., Rouillard, A. P., & Davies, J. A. 2017b, ApJL, 835, L7
Sundberg, T., Burgess, D., Scholer, M., Masters, A., & Sulaiman, A. H. 2017, ApJL, 836, L4
Vech, D., & Malaspina, D. M. 2021, JGRA, 126, e29567
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Wilcoxon, F. 1947, Biometrics, 3, 119
Zewdie, G. K., Valladares, C., Cohen, M. B., et al. 2021, SpWea, 19, e2020SW002639
Zhao, L.-L., Zank, G. P., Adhikari, L., et al. 2020, ApJS, 246, 26
Zheng, J., Hu, Q., Chen, Y., & le Roux, J. 2017, JPhCS, 900, 012024