Automatic Detection of Interplanetary Coronal Mass Ejections in Solar Wind In Situ Data

H. T. Rüdisser[1,2], A. Windisch[1,3,4,5], U. V. Amerstorfer[6], C. Möstl[6], T. Amerstorfer[6], R. L. Bailey[7], and M.A. Reiss[6]

[1]Know-Center GmbH, Inffeldgasse 13, 8010 Graz, Austria
[2]Institute of Physics, University of Graz, Universitätsplatz 5, 8010 Graz, Austria
[3]Institute of Interactive Systems and Data Science, Graz University of Technology, Inffeldgasse 13, 8010 Graz, Austria
[4]Department of Physics, Washington University in St. Louis, MO 63130, USA
[5]RL Community, AI AUSTRIA, Wollzeile 24/12, 1010 Vienna, Austria
[6]Space Research Institute, Austrian Academy of Sciences, Schmiedlstraße 6, 8042 Graz, Austria
[7]Zentralanstalt für Meteorologie und Geodynamik, Hohe Warte 38, 1190 Vienna, Austria

Key Points:

- We automatically detect interplanetary coronal mass ejections in solar wind in situ data.
- We achieve a True Skill Statistic (TSS) of 0.64, Recall of 0.67 and Precision of 0.70 on data from Wind.
- We propose a pipeline generally applicable to time series event detection problems.

Corresponding author: Hannah Rüdisser, hruedisser@know-center.at

Abstract

Interplanetary coronal mass ejections (ICMEs) are one of the main drivers for space weather disturbances. In the past, different approaches have been used to automatically detect events in existing time series resulting from solar wind in situ observations. However, accurate and fast detection still remains a challenge when facing the large amount of data from different instruments. For the automatic detection of ICMEs we propose a pipeline using a method that has recently proven successful in medical image segmentation. Comparing it to an existing method, we find that while achieving similar results, our model outperforms the baseline regarding training time by a factor of approximately 20, thus making it more applicable for other datasets. The method has been tested on in situ data from the Wind spacecraft between 1997 and 2015 with a True Skill Statistic (TSS) of 0.64. Out of the 640 ICMEs, 466 were detected correctly by our algorithm, producing a total of 254 False Positives. Additionally, it produced reasonable results on datasets with fewer features and smaller training sets from Wind, STEREO-A and STEREO-B with True Skill Statistics of 0.56, 0.57 and 0.53, respectively. Our pipeline manages to find the start of an ICME with a mean absolute error (MAE) of around 2 hours and 56 minutes, and the end time with a MAE of 3 hours and 20 minutes. The relatively fast training allows straightforward tuning of hyperparameters and could therefore easily be used to detect other structures and phenomena in solar wind data, such as corotating interaction regions.

Plain Language Summary

Interplanetary coronal mass ejections (ICMEs) are part of space weather and can have severe effects on human technology. Since the detection of these events is often difficult, experts are usually needed to detect them in time series from different spacecraft. Even though, there have been attempts to solve this problem using machine learning, it is far from being solved. We propose a machine learning model, that manages to find the start and end times with improved accuracy and takes less time to train. Even though our results are quite promising, it is important to notice that there are many different catalogs documenting ICMEs, that do not always agree. Therefore, there is no unique solution to this problem, which makes it difficult for a machine learning method to learn the features of an ICME. Nevertheless, our model can be expected to make a substantial contribution to the area of space weather forecast in the future.

1 Introduction

Interplanetary coronal mass ejections (ICMEs) are the interplanetary counterpart of coronal mass ejections (CMEs). Continually interacting with planetary environments, these impactful manifestations of solar activity drive the most extreme forms of space weather in our solar system and are not yet fully understood. Nevertheless, their geoeffectiveness and capability to trigger magnetic storms have societal impacts, which cannot be disregarded. Automatically detecting ICMEs in solar wind in situ data is a crucial part for accurately forecasting these events and their consequences.

Initial studies (Gosling et al., 1973; Burlaga et al., 1981; Klein & Burlaga, 1982) and consecutive investigations of in situ measurements (Kilpua et al., 2017) led to a more thorough understanding of the typical appearance of ICMEs. Figure 1 depicts an example event exhibiting the typical signatures of an ICME featuring a magnetic cloud (MC), notwithstanding the existence of events without the presence of aforementioned structure (Rouillard, 2011). The top panel shows a smooth rotation of the magnetic field components and an enhanced total magnetic field compared to the surrounding ambient solar wind. The second panel shows a first enhanced, but then monotonically declining speed profile compared to the preceding solar wind. Both the velocity and the magnetic field exhibit a sudden sharp jump, the so called shock, which is followed by the sheath, a tur-

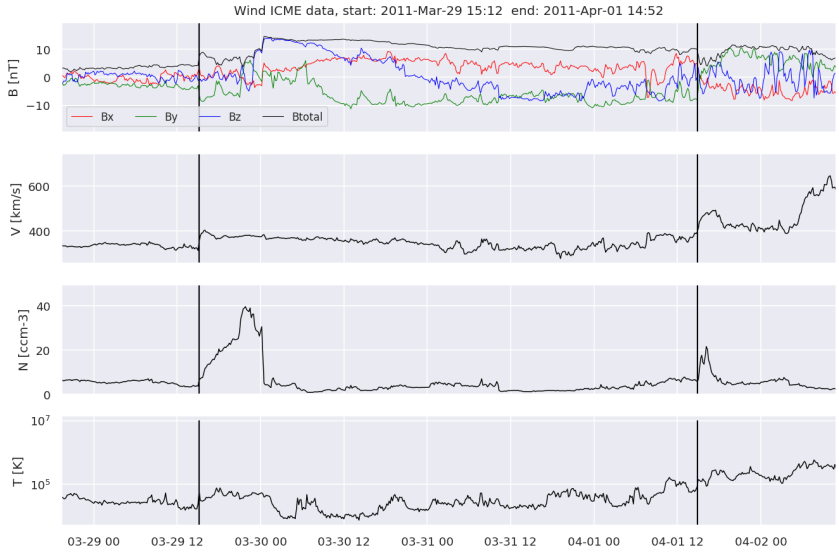


Figure 1. Solar wind in situ data from the Wind spacecraft located at the Lagrangian Point L1, showing an ICME. The solid vertical lines delimitate the event, including shock, sheath and magnetic cloud. From top to bottom: magnetic field amplitude and components, solar wind velocity, proton density, proton temperature.

bulent region preceding the magnetic cloud. The third panel depicts an extreme proton density decrease after the proton density increase during the shock and sheath, and the fourth panel shows a reduced proton temperature.

These standard criteria are generally used for the identification of ICMEs, complemented by several other features (Zurbuchen & Richardson, 2006; Chi et al., 2016; Kilpua et al., 2017). However, the circumstance that not all ICMEs even exhibit standard features and that there exists no feature present in all ICMEs hinders a standardized identification method and fuels the need for time consuming visual expert labeling. Nevertheless, this visual inspection is highly biased, emphasized through the existence of several catalogs. Even though they have a significant overlap agreeing on most of the events, some use more loose criteria than others resulting in a higher number of ICMEs. Additionally, the definition of start and end time varies from one expert to another (Lepping et al., 2006; Jian et al., 2006; Richardson & Cane, 2010; Chi et al., 2016; Nieves-Chinchilla et al., 2018; Nguyen et al., 2019; Möstl et al., 2020).

Nguyen et al. (2019) highlighted the need of an automatic detection and proposed a pipeline based on deep learning, which produced impressive results. During 1997 − 2015, 465 out of 640 ICMEs were found, while producing a total of 215 False Positives. However, the training duration of around 35 hours and the deviation of predicted start and end time from the ground truth leave room for improvement. The start of the ICME could be predicted with a mean absolute error (MAE) of around 4 hours and 4 minutes, while the predicted end time deviated from the label by a mean absolute error of 4 hours and 46 minutes. The study by Nguyen et al. (2019) was, however, one of the first to address this topic, and one in a growing number of studies on the application of machine learning in the space sciences. The prediction of the B_z component from upstream in situ observations (Reiss et al., 2021), forecasting global geomagnetic activity (Topliff et al., 2020) or the timing of the solar wind propagation delay between the Lagrangian Point L1 and Earth (Baumann & McCloskey, 2021) are just a few examples of how machine learning is being applied in the area of space weather.

The automatic detection of ICMEs is only one specific instance of the widely known challenge of detecting events in time series. Approaches tackling this find application in many other fields, such as the identification of sleep arousals in EEG (Electroencephalogram) data (Li & Guan, 2021) or human activity recognition (Yang et al., 2015). Even the solar wind in situ data itself features many more phenomena beyond ICMEs, for instance corotating interaction regions (CIRs).

A fairly recent breakthrough in the area of machine learning was the U-Net (Ronneberger et al., 2015) and subsequent related advances. Its original application was the segmentation of biomedical images, where spatial information is as valuable as feature information. The problem of detecting ICMEs resembles the initial implementation in this aspect and therefore suggests the investigation of segmentation algorithms for this specific use case. Furthermore, segmentation algorithms have been shown to successfully deal with other time series event detection problems, such as sleep staging (Perslev et al., 2019).

In this article, we rephrase the automatic detection of ICMEs as a time series segmentation problem and propose a pipeline using a variation of the ResUNet++ (Jha et al., 2019), which shows improvements compared to an existing method. A comparable study was published during the finalization of this manuscript by Chen et al. (2022). Nevertheless, we use a different validation method focusing on generalization, as conventional within the machine learning community to avoid overfitting and predict the performance on new unseen data more accurately. Furthermore, our pipeline was tested on additional data from STEREO-A and STEREO-B. Finally, we provide insight on the MAE made for the prediction of start and end times to explore the suitability of our method for the area of real-time detection. The general concept of our pipeline was first presented at the Europlanet Science Congress 2021 (Rüdisser & Amerstorfer, 2021). In Section 2, we present the data used for the training, validation and testing of our pipeline. In Section 3, we give a detailed description of our pipeline including preprocessing, training and postprocessing and present the validation methods used for the evaluation of the performance. In Section 4 we summarize the results and compare it to the results of Nguyen et al. (2019). Finally, Section 5 discusses the results of this article and provides a short outlook on possible improvements and applications in the future.

2 Data

2.1 In situ Data

To compare the performance of our pipeline to the method proposed in Nguyen et al. (2019), we use the same data and catalog. The 30 primary features were provided by the Magnetic Field Investigation (Lepping et al., 1995), Solar Wind Experiment (Ogilvie et al., 1995), and 3D Plasma and Energetic Particles Experiment (Lin et al., 1995) on board the spacecraft Wind (Wilson et al., 2021) between 1997 October 1 and 2016 January 1 and three additional features were computed, in accordance to Nguyen et al. (2019). Additionally, we tested the pipeline on STEREO-A and STEREO-B data, which include fewer features than the one used by Nguyen et al. (2019). These two datasets were complemented with a Wind dataset, reduced to the same features. The features for each dataset are summarized in Table 1.

2.2 Catalogs

We used the ICME catalog by Nguyen et al. (2019), which consists of different ICME lists (Jian et al., 2006; Lepping et al., 2006; Richardson & Cane, 2010; Chi et al., 2016; Nieves-Chinchilla et al., 2018), plus the ICMEs that were detected by the pipeline and thereupon added to the catalog. Thus, the total number of ICMEs in this specific catalog is much higher, as can be seen in Figure 2, where the number of events per year listed in a particular catalog is compared.

Table 1. Features in the full Wind dataset as used by Nguyen et al. (2019) and the reduced Wind, STEREO-A and STEREO-B datasets. • indicates the presence of a parameter and ○ indicates that the parameter had to be computed during preprocessing.

Parameter	Description	Full Dataset	Reduced Dataset
V	Bulk velocity	•	•
V_x,V_y,V_z	bulk velocity components	•	
B, B_x, B_y, B_z,	magnetic field components	•	•
σ_{B_x},σ_{B_y},σ_{B_z}	root mean square of magnetic field components	•	
N_p	proton density (from moment)	•	•
$N_{p,nl}$, $N_{a,nl}$	proton density and α particle density (from non-linear analysis)	•	
$\Phi_1, \ldots, \Phi_{15}$	15 canals of proton flux	•	
T_p	proton temperature		•
β	ratio between thermal and magnetic pressure	○	○
P_{dyn}	dynamic pressure	○	○
σ_B	normalized magnetic fluctuations	○	
T_p/T_{ex}	ratio between proton temperature and expected temperature		○

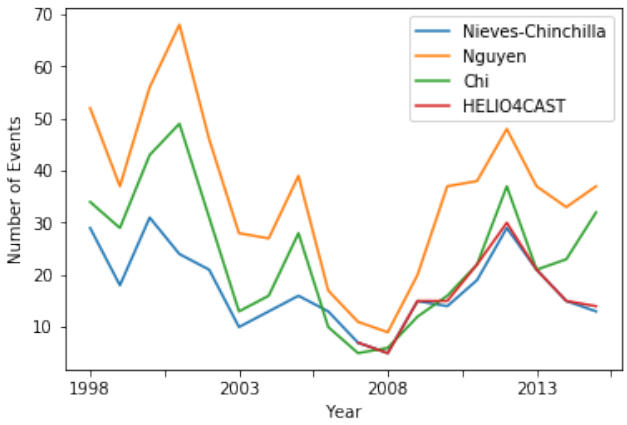


Figure 2. Comparison of the number of events listed in different catalogs: Chi et al. (2016), Nieves-Chinchilla et al. (2018), Nguyen et al. (2019) and HELIO4CAST (Möstl et al., 2020).

For the STEREO-A and STEREO-B datasets, as well as the reduced Wind dataset, we used the HELIO4CAST ICMECAT (Möstl et al., 2020, 2022). Guided by the criteria in Nieves-Chinchilla et al. (2018), it contains only events that show clear signatures of magnetic structures, called magnetic obstacles.

3 Machine Learning and Pipeline

In the following, we present our pipeline for the automatic detection of ICMEs.

3.1 Preprocessing

During preprocessing, additional features are computed according to Section 2.1. To eliminate missing data points, the dataset was resampled to a 10 minute resolution, which is sufficient for ICMEs that have average durations of about one day at 1 AU. Nevertheless, it would be possible to choose a higher resolution if demanded by the scale of the structures that need to be segmented. Since there exist significant differences in orders of magnitude among the various features, they all were normalized and scaled to have an average of 0 and a standard deviation of 1.

We use a sliding window approach with a window size equal to 1024 and a stride of 120 to preprocess the multivariate time series. Thereby, we extract samples covering approximately one week in time and thus include even the longest events as a whole. The labels are one-dimensional segmentation maps consisting of the values 0 or 1 for each point in time with a resolution of 10 minutes, indicating whether an ICME is taking place or not. It is important to notice that we do not use overlapping windows for the test set in order to avoid redundancies.

3.2 Architecture of the Model

The backbone of our model is a variation of a U-Net (Ronneberger et al., 2015), more precisely a Deep Residual U-Net, as proposed in Zhang et al. (2018). In general, a U-Net consists of a contracting path and an expansive path, arranged in a U-shaped architecture. The contraction path gradually reduces spatial information, while feature information is increased. The missing spatial information is then reintroduced through concatenations of up-convolutions and high-resolution features stemming from the contracting path. While the original U-Net uses simple convolutional blocks, the Deep Residual U-Net makes use of residual units instead (He et al., 2015), in order to overcome the degradation problem and further improve semantic segmentation performance.

Jha et al. (2019) proposed the ResUNet++, which is an improved ResUNet architecture for colonoscopic image segmentation and significantly outperforms other state of the art algorithms used for semantic segmentation. The ResUNet architecture is extended through the application of squeeze-and-excitation units (Hu et al., 2019), atrous spatial pyramid pooling (ASPP; He et al., 2014; Chen et al., 2017, 2018) and attention units (Vaswani et al., 2017), thus enhancing the focus on relevant features and areas, increasing generalization, reducing computational cost and capturing channel-wise dependencies.

In this paper, we adapt the ResUNet++ to create an automatic event detection based on time series segmentation. We expand the dimensions of our input windows and thereby generate input "images" of size $(t, 1, C)$, which are subsequently fed into the network. Here, t is the time domain and equal to 1024, as mentioned in Section 3.1, C depicts the numbers of channels, in other words, the features present in our dataset (33 for Wind, 11 for STEREO-A, STEREO-B and the reduced Wind dataset).

The network itself consists of three parts: Encoding, Bridge and Decoding. A block diagram of the used model architecture is shown in Figure 3 and a detailed description of the parameters and output sizes can be found in Table 2.

The encoding path, which can be seen on the left in Figure 3, consists of one stem block and two residual units. The stem block uses batch normalization and a Rectified Linear Unit (ReLU) combined with two 3×3 convolutional layers. Its output is passed through a squeeze-and-excitation block. The two residual units are each a combination of batch normalization, ReLU activation and two successive 3×3 convolutional layers. The first of these is in fact a strided convolution, applied to reduce the spatial dimension of the feature maps to a quarter. The identity mapping connects the input and output of the encoder block. Each residual unit is once again followed by a squeeze-and-excitation block.

The bridge consists of an ASPP block, enlarging the field-of-view of the filters coming from the encoder. Similar to the encoder, the decoding path consists of residual units, followed by squeeze-and-excitation units, as well, but uses an attention block beforehand to increase the effectiveness of feature maps. A nearest-neighbour up-sampling of the lower level feature maps is then concatenated with feature maps from their corresponding part in the encoder path.

Ultimately, the last level of the decoder is followed by ASPP and a 1×1 convolution with sigmoid activation in order to project the multi-channel feature maps into the final segmentation map.

3.3 Training

We use the Adam optimizer with a cyclical learning rate (lr), as introduced in Smith (2017), with $lr_{base} = 0.00001$, $lr_{max} = 0.01$, and $stepsize = 1000$. Overfitting is reduced through EarlyStopping by monitoring the validation loss and stopping if no improvement has been made for 50 epochs. Dice Loss with a smoothing factor of 1 is used as loss function in order to account for the class imbalance (Jadon, 2020). The training examples are shuffled after each epoch and we use a batch size of 32.

3.4 Postprocessing

During postprocessing, the previously continuous label predicted by the algorithm is converted to a binary label, by applying a threshold of 0.5. The following step is the generation of an ICME catalog by extraction of areas of label 1. Furthermore, events shorter than 3 hours are removed from the catalog, since only around 0.3% of the events in Nguyen et al. (2019) fall into this category. Thus, these predicted events are considered an inconsistent prediction.

3.5 Validation

In general, it is important to keep in mind that a supervised machine learning algorithm learns from a ground truth, which may already be biased. Furthermore, the catalogs used for creating the ground truth may not be exhaustive. For the evaluation of our pipeline we compare the ground truth to our prediction. However, we define two different validation methods, which are presented hereafter.

One way of evaluating a machine learning model is the straightforward comparison of the ground truth label and the predicted label (pointwise validation). Each point in time is then sorted into one of the categories as shown in Table 3. Using these categories, one can compute the classification metrics listed in Table 4.

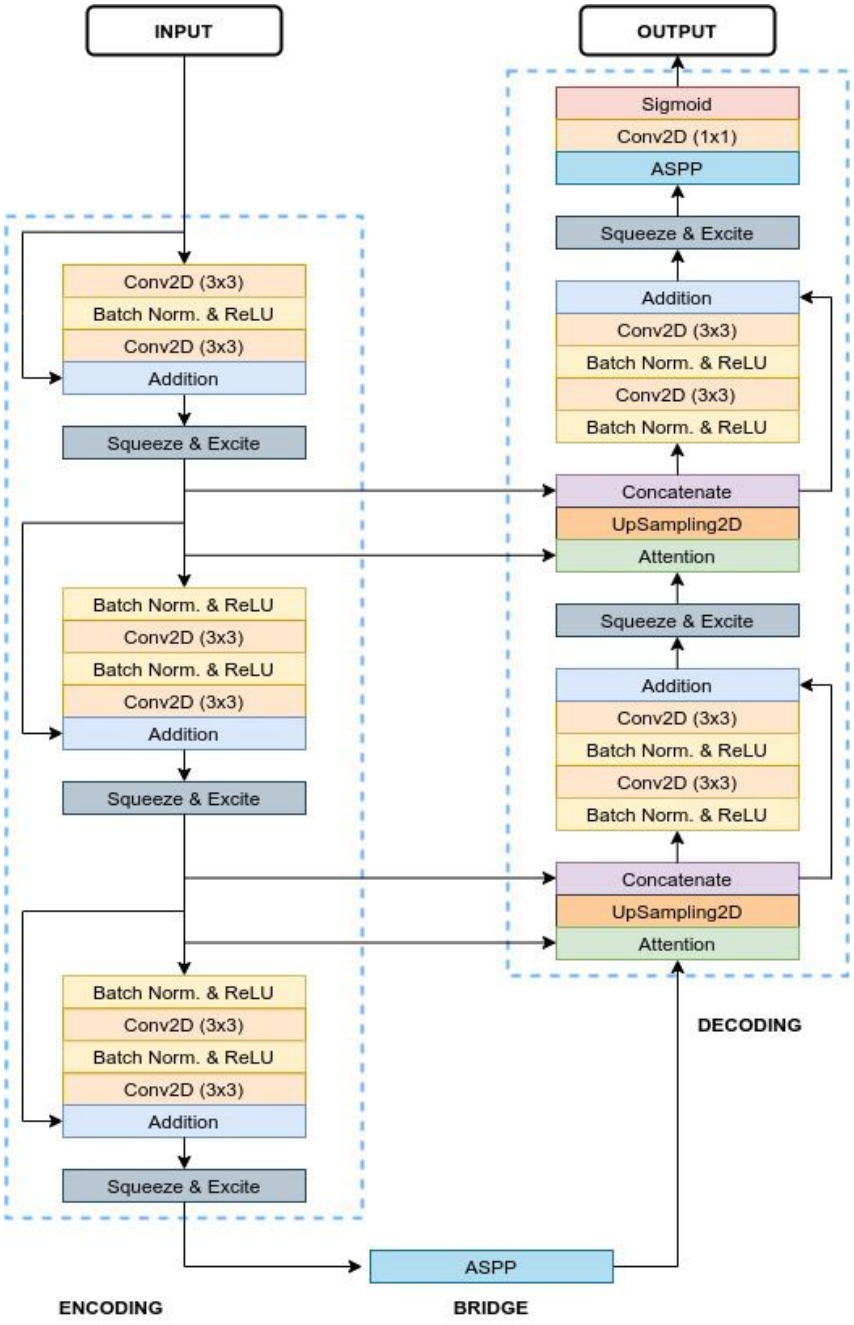


Figure 3. Block diagram of the model architecture.

Table 2. Model Architecture.

	Block	Layer	Filter	Stride/Size	Output Size
Input					1024 x 1 x 33
Encoding	Stem Block	Conv2D	3x3/64	1	1024 x 1 x 64
		Batch Normalization & ReLU			1024 x 1 x 64
		Conv2D	3x3/64	1	1024 x 1 x 64
	Squeeze & Excitation				1024 x 1 x 64
	Residual Unit	Batch Normalization & ReLU			1024 x 1 x 64
		Conv2D	3x3/128	4	256 x 1 x 128
		Batch Normalization & ReLU			256 x 1 x 128
		Conv2D	3x3/128	1	256 x 1 x 128
	Squeeze & Excitation				265 x 1 x 128
	Residual Unit	Batch Normalization & ReLU			256 x 1 x 128
		Conv2D	3x3/256	4	64 x 1 x 256
		Batch Normalization & ReLU			64 x 1 x 256
		Conv2D	3x3/256	1	64 x 1 x 256
	Squeeze & Excitation				64 x 1 x 256
Bridge	Atrous Spatial Pooling				64 x 1 x 512
Decoding	Attention				64 x 1 x 512
	Connection	Upsampling2D		(4,1)	256 x 1 x 512
		Concatenate			256 x 1 x 640
	Residual Unit	Batch Normalization & ReLU			256 x 1 x 640
		Conv2D	3x3/256	1	256 x 1 x 256
		Batch Normalization & ReLU			256 x 1 x 256
		Conv2D	3x3/256	1	256 x 1 x 256
	Squeeze & Excitation				256 x 1 x 256
	Connection	Upsampling2D		(4,1)	1024 x 1 x 256
		Concatenate			1024 x 1 x 320
	Residual Unit	Batch Normalization & ReLU			1024 x 1 x 320
		Conv2D	3x3/128	1	1024 x 1 x 128
		Batch Normalization & ReLU			1024 x 1 x 128
		Conv2D	3x3/128	1	1024 x 1 x 128
	Squeeze & Excitation				1024 x 1 x 128
	Atrous Spatial Pooling				1024 x 1 x 64
	Segmentation	Conv2D	1x1/1		1024 x 1 x 1
		Sigmoid			1024 x 1 x 1

Table 3. Standard Confusion Matrix explaining the four different categories.

		Predicted Label	
		0	**1**
True Label	**0**	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Table 4. Overview of pointwise classification metrics.

Metric	**Definition**
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
Dice Coefficient	$\frac{2TP}{2TP+FP+FN}$
True Skill Statistics	$\frac{TP}{TP+FN} + \frac{TN}{FP+TN}$ - 1
Intersection Over Union	$\frac{TP}{TP+FN+FP}$

Another way of evaluating a machine learning model is the comparison of the ground truth catalog and the predicted catalog (event based validation), which was generated during the postprocessing step (Nguyen et al., 2019). A predicted event is counted as a TP if the overlap with a ground truth event is at least 10%. An FP is an event for which this does not hold, and an FN is a ground truth event for which no TP can be found. Considering the fact that no TN can be defined, Precision and Recall serve as evaluation metrics and are computed following Table 4. Recall, also known as sensitivity, is a measure of the fraction of instances correctly predicted to be positive (True Positives) among all positive instances. Precision, also known as positive predictive value, describes the fraction of True Positives among all instances predicted to be positive. Since a higher recall often comes with a lower precision and vice versa, Dice Coefficient, True Skill Statistics (TSS) and Intersection Over Union (IoU) are metrics which attempt to measure the overall performance. Even though Dice Coefficient and IoU are quite similar, when averaging over all examples, IoU tends to penalize single incorrectly classified examples more than the Dice Coefficient.

The severe class imbalance in our problem demands a profound and in-depth analysis of the results obtained. Since ICMEs are rare and the solar wind is undisturbed most of the time, predicting the background label at all times and completely disregarding the existence of events would already produce seemingly acceptable results in terms of accuracy. While this can be avoided using other evaluation metrics as shown in Table 4, there is a high variation in events per year, as depicted in Figure 2. Therefore, there is a non negligible chance, that either test or validation set might contain considerably less events than the training set or vice versa. In general, cross validation is commonly used to prevent the outcome to be more optimistic or pessimistic than the actual performance of the model, and therefore misleading. However, in the case of a highly skewed distribution, adaptions of the cross validation technique are required. More precisely, the distribution of different classes has to stay approximately equal over the training set, validation set and test set. Nevertheless, random splitting with a uniform probability distribution leads to a high imbalance in our case. Thus, we use manual stratification to enforce the class distribution in each split of the data to match the overall distribution. To be exact, we split the data by year and divide it into subsets consisting of three years, such that the number of ICME events is approximately equal in each subset. Subsequently, we perform classic 6-fold cross validation. On average, we use around 107 events for val-

Table 5. Overview of the different balanced splits on the complete Wind dataset between 1998 and 2015 used for crossvalidation.

	Part 1 **2000,2005,2008**	**Part 2** **1998,2007,2009**	**Part 3** **2003,2004,2012**	**Part 4** **2001,2010,2014**	**Part 5** **1999,2002,2006**	**Part 6** **2011,2013,2015**
1	Test	Val	Train			
2		Test	Val	Train		
3			Test	Val		Train
4	Train			Test	Val	
5		Train			Test	Val
6	Val	Train				Test

Table 6. Overview of the different balanced splits on the reduced datasets between 2007 and 2019 used for cross validation.

	Part 1	**Part 2**	**Part 3**	**Part 4**
Wind:	2008,2012,2016	2007,2009,2011	2013,2014,2018	2010,2015,2017,2019
STEREO A:	2009,2013,2015	2008,2010,2014,2018	2017,2012,2017	2011,2016,2019
STEREO B:	2007,2012	2008,2013	2009,2011	2010,2014
1	Test	Val	Train	
2		Test	Val	Train
3	Train		Test	Val
4	Val	Train		Test

idation and testing and around four times as many for training. This process is visualized in Table 5. The same procedure is carried out on the reduced datasets, although the various subsets vary in size and only add up to four different splits. The parts specified in Table 6 contain 44 events each on average for the reduced Wind dataset, 50 for STEREO-A and 44 for STEREO-B.

4 Results

4.1 Full Wind Dataset

The pointwise results of our pipeline for the full Wind dataset are summarized in Table 7. A confusion matrix visualizing the results is depicted in Figure 4 and a comparison of the ground truth label and the predicted label for a short time period can be seen in Figure 5. While the pipeline proposed by Nguyen et al. (2019) shows a higher recall, our pipeline achieves a higher precision. Since varying the threshold for the pipeline of Nguyen et al. (2019) increases one of aforementioned metrics at the expense of the other, the results can be expected to be quite similar. This assumption proves true, when analysing other performance metrics. While our pipeline achieves a slightly higher Dice Coefficient and IoU, the TSS is higher for the pipeline of Nguyen et al. (2019).

Performing the event based validation, both pipelines find an approximately equal number of events, while our pipeline produces a higher number of false positives, result-

Table 7. Pointwise results for the full Wind dataset, averaged over all splits for our pipeline as well as the pipeline proposed by Nguyen et al. (2019).

Metric	**Our Pipeline**	**Nguyen**
Recall	0.67	0.79
Precision	0.70	0.56
Dice Coefficient	0.69	0.66
True Skill Statistic	0.64	0.72
IoU	0.52	0.49

Table 8. Event based results for the full Wind dataset, averaged over all splits for our pipeline as well as the pipeline proposed by Nguyen et al. (2019).

Metric	**Our Pipeline**	**Nguyen**
Recall	0.73	0.72
Precision	0.65	0.68
Dice Coefficient	0.69	0.70
True Positives	466	465
False Negatives	174	175
False Positives	254	215
MAE Start Time	2h56m	4h4m
MAE End Time	3h20m	4h46m

Figure 4. Confusion matrix visualizing the results of the full Wind dataset.

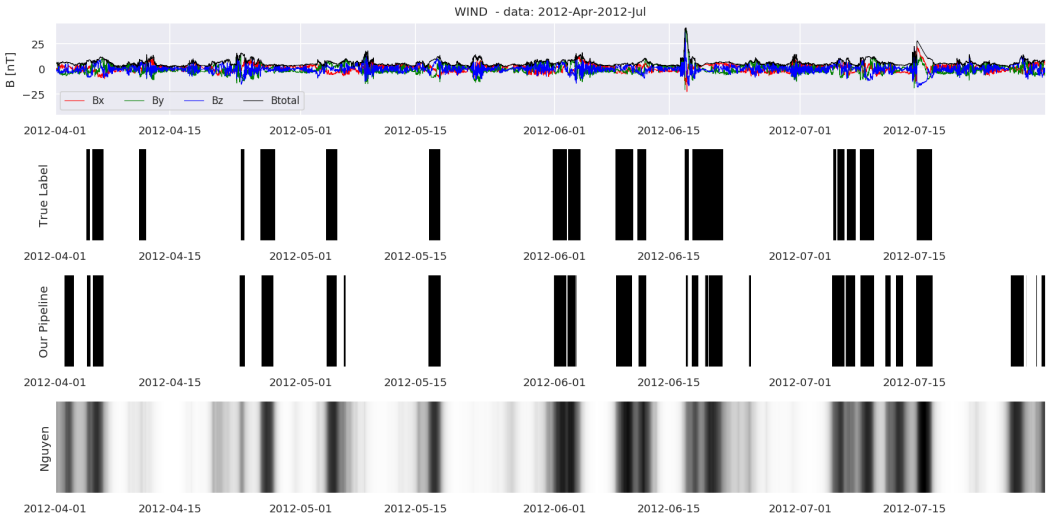


Figure 5. Comparison of the ground truth label and the predicted labels for the full Wind dataset in the time period of April 2012 to July 2012. The top panel shows the magnetic field and its components, the second panel shows the ground truth label, the third panel depicts the label predicted by our pipeline, and the label predicted by Nguyen et al. (2019) is shown in the bottom panel.

ing in a lower Precision. Nevertheless, the start and end time can be predicted more accurately by our pipeline, with mean average errors of over an hour less than for the pipeline of Nguyen et al. (2019). Not surprisingly, the prediction of start times is more accurate in general than for the end times, since the presence of a shock in some events tends to indicate the start time more clearly.

4.2 Reduced Wind, STEREO-A and STEREO-B Datasets

Table 9 summarizes the results for each of the reduced datasets. At first, three models were trained on only one of the datasets each and are averaged over all splits. In a second run, one model was trained on all three datasets simultaneously. The event based results for the reduced datasets can be seen in Table 10. As expected, the results for the full Wind dataset are significantly better, than for the reduced sets. Nevertheless, the performance on the STEREO-A dataset seems to stand out among the other reduced datasets. While the pointwise results on all three datasets are relatively similar and slightly improve for the joint training, the event based evaluation shows more considerable differences. The Dice Coefficient for the separate training is significantly higher for the STEREO-A sets, but does not improve for the joint training in contrast to the other datasets. Additionally, the prediction of the start time for the STEREO-A dataset, when trained separately, is even more accurate than for the full Wind dataset.

5 Discussion

To allow the community to compare future studies with our findings, the source code, ICMECATv2.0 catalog, and related data are available online as outlined in Section 7.

Figure 5 graphically illustrates the way in which our predictions work. Visual inspection already suggests the high level of accordance between the predicted label and

Table 9. Pointwise results for the reduced datasets, averaged over all splits.

	Separate Training			Joint Training		
Metric	**Wind**	**STEREO A**	**STEREO B**	**Wind**	**STEREO A**	**STEREO B**
Recall	0.55	0.54	0.57	0.58	0.58	0.55
Precision	0.62	0.67	0.58	0.63	0.68	0.68
Dice Coefficient	0.58	0.60	0.57	0.60	0.63	0.60
True Skill Statistic	0.53	0.52	0.54	0.56	0.57	0.53
IoU	0.41	0.43	0.40	0.43	0.46	0.43

Table 10. Event based results for the reduced datasets, averaged over all splits.

	Separate Training			Joint Training		
Metric	**Wind**	**STEREO A**	**STEREO B**	**Wind**	**STEREO A**	**STEREO B**
Recall	0.58	0.66	0.70	0.60	0.66	0.63
Precision	0.50	0.62	0.52	0.52	0.60	0.63
Dice Coefficient	0.54	0.64	0.60	0.56	0.63	0.63
True Positives	102	133	104	107	134	93
False Negatives	75	70	44	70	69	55
False Positives	101	82	95	100	90	54
MAE Start Time	3h40m	2h31m	3h3m	3h42m	3h14m	2h54m
MAE End Time	4h13m	4h56m	5h13m	4h8m	4h29m	4h32m

the ground truth label. Even though not every single event is captured in its entirety, the algorithm definitely highlights areas of interest.

Not surprisingly, the results for the full Wind dataset are significantly better than the results for the reduced datasets, which contain fewer parameters. Nevertheless, a considerable number of ICMEs have been detected for these datasets as well. Interestingly, only a minor improvement could be gained by combining the datasets during training. In deep learning, more data usually leads to better generalisation and thereby better results. Even though the Dice Coefficient slightly increases, the underlying datasets and their respective labels seem to contain certain inconsistencies. This assumption is also supported by the fact that the performance on the STEREO-A dataset is considerably better than on the other two reduced datasets. A possible explanation lies in the ambiguity included in the labeling. The differences shown from one catalog to another as well as the amount of events retrospectively added through the pipeline of Nguyen et al. (2019) display the disagreement between experts. These biases are adopted as the algorithm learns from the labels and result in the fact that a certain threshold regarding performance can probably not be exceeded.

As highlighted in Nguyen et al. (2019), there is a substantial chance of false positives actually being ICMEs or at least ICME-like structures. Figure 6 shows an example of a false positive that resembles an ICME. In comparison, Figure 7 shows a false negative, where the signatures are not as prominent as in other examples. To sum up, the incompleteness of the catalogs impairs training and evaluation at the same time.

Nguyen et al. (2019) proposed a deep learning approach predicting a similarity measure for the automatic detection of ICMEs. Throughout this study, we conducted a different strategy using a single but more complex architecture for the segmentation of time series. Creating a 2D similarity map has the advantage of switching between high Precision and high Recall by simply adjusting the threshold, since the algorithm has learned

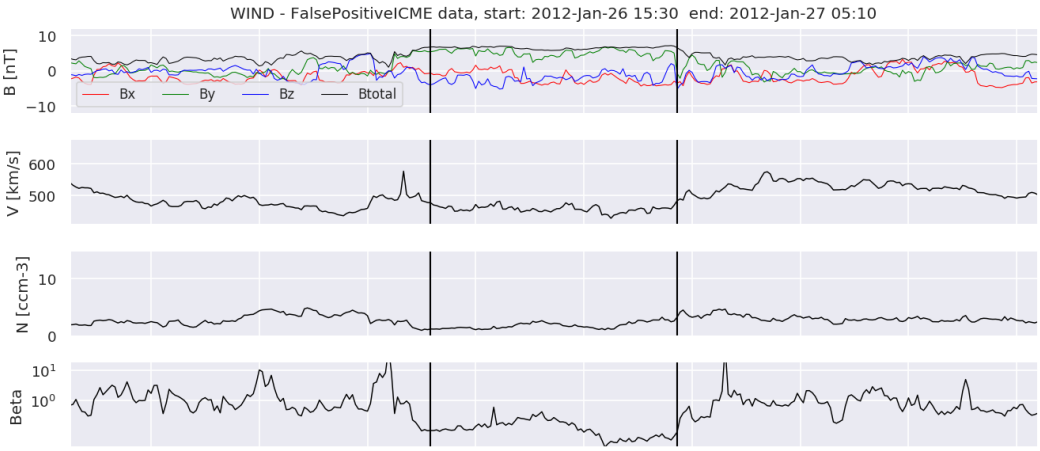


Figure 6. Example of a false positive found by the algorithm.

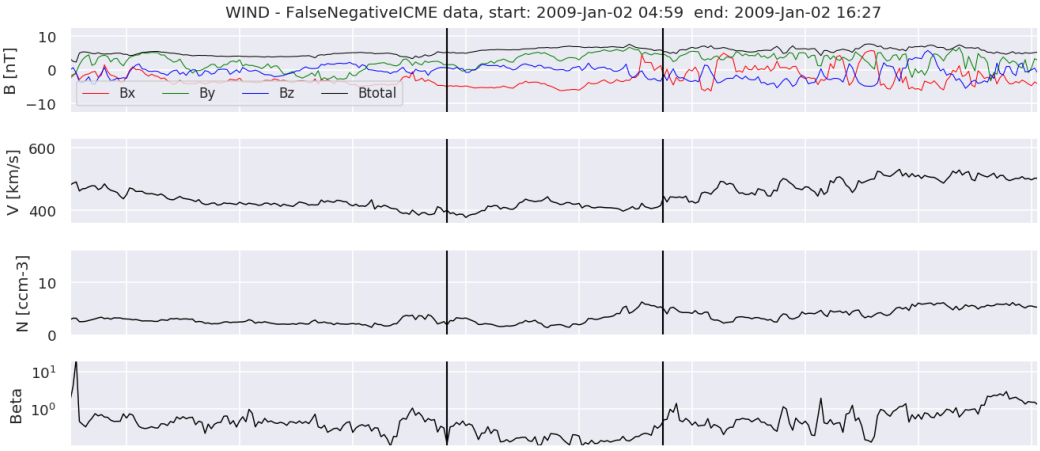
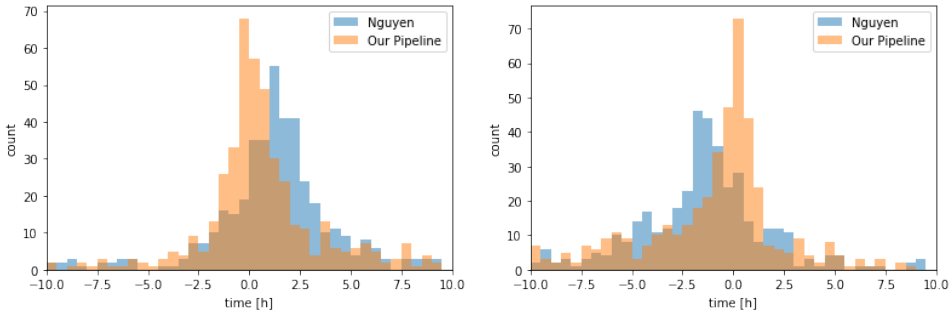


Figure 7. Example of a false negative not found by the algorithm.



(a) Deviation from start time.

(b) Deviation from end time.

Figure 8. Histogram showing the deviation of the predicted start from the ground truth start in hours for Nguyen ($MAE = 4h4m$) and our pipeline ($MAE = 2h56m$), and the deviation of the predicted end from the ground truth end in hours for Nguyen ($MAE = 4h46m$) and our pipeline ($MAE = 3h20m$).

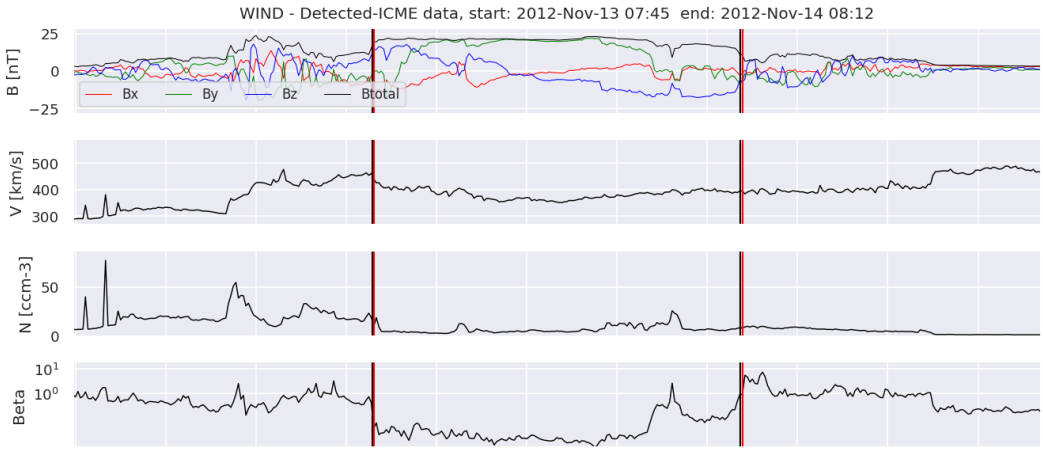


Figure 9. Example of a detected ICME along with its ground truth (black) and predicted limits (red).

to predict values between 0 and 1. On the contrary, our pipeline highly favors an output of either 0 or 1, thus forfeiting the option of being interpreted as a probability for the occurrence of an ICME. Nevertheless, our pipeline shows promising results. Training and testing times were reduced by a factor of approximately 20 compared to Nguyen et al. (2019), while simultaneously achieving reasonable Recall and Precision. Additionally, the binary segmentation has the advantage of relatively high agreement between predicted start and end time and the ground truth limits, compared to the similarity approach, which can be seen in Figure 8. Our pipeline manages to find the start of an ICME with a mean absolute error of around 2 hours and 56 minutes, and the end time with a mean absolute error of 3 hours and 20 minutes, which is more than an hour less than achieved by Nguyen et al. (2019). Examples showing the ground truth as well as the predicted limits can be found in Figure 9, 10 and 11.

To focus more on either high Recall or high Precision, one would have to adjust the loss function to suit specific needs. Another possibility to evade this drawback would be to allow the model to learn continuous values as well, by adjusting the labels accord-

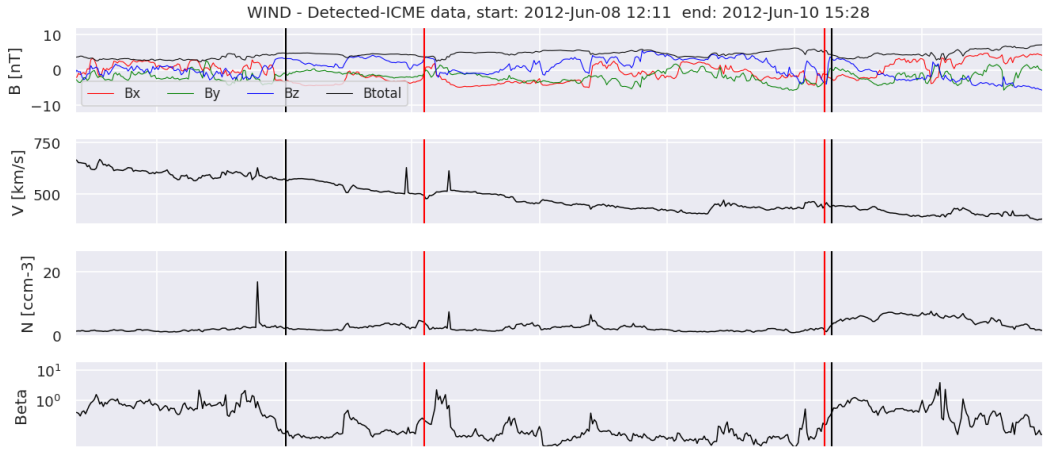


Figure 10. Example of a detected ICME along with its ground truth (black) and predicted limits (red).

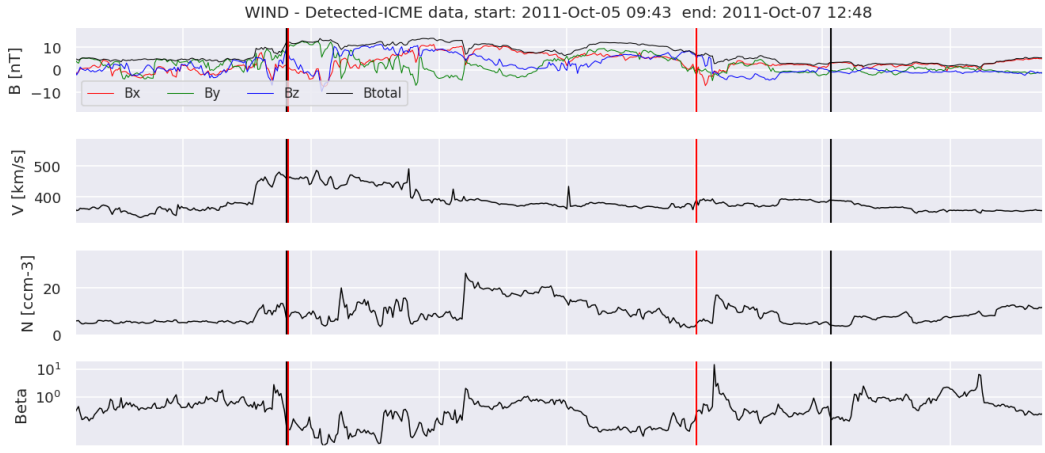


Figure 11. Example of a detected ICME along with its ground truth (black) and predicted limits (red).

ingly, or pursue alternative approaches, such as probabilistic machine learning. However, these options fall beyond the scope of this work and have not been assessed yet.

Furthermore, the straightforward implementation of our algorithm paves the way for a simple extension to multiclass segmentation. This could be applied to simultaneously detect ICMEs and corotating interaction regions (CIRs) as well as other structures within the solar wind. Additionally, the application of detecting only the shock and sheath of an event is a presumably easier task and may conduce to other tools, using this information as input.

Finally, it is important to recognize that, given the complexity of the problem, there is no unique solution to it. Additionally, an algorithm using deep learning can only ever be as good as the label it is learning from and thus is expected to be biased by some degree of uncertainty.

To sum up, our pipeline serves as a fast and relatively reliable tool for the automatic segmentation of time series. Considering the difficulties included in the specific task of detecting ICMEs, a considerable improvement regarding the detection rate can presumably not be expected through changes in the architecture of the model. However, our pipeline shows significant gains concerning the prediction of start and end time of detected events, compared to the pipeline by Nguyen et al. (2019). In the future, the exploration of probabilistic machine learning might substantially improve the output of our pipeline considering its application as an early warning system. Nevertheless, its combination with other forecasting tools can already be expected to serve as a significant contribution in the area of space weather forecast.

Our pipeline is comparable to the one proposed by Chen et al. (2022). Nevertheless, we use a different validation method focusing on generalization, as conventional within the machine learning community to avoid overfitting and predict the performance on new unseen data more accurately. Furthermore, our pipeline was tested on additional data from STEREO-A and STEREO-B. Finally, we provide insight on the MAE made for the prediction of start and end times to explore the suitability of our method for the area of real-time detection.

6 Summary

The ever-growing amount of solar wind in situ data demands a fast, reliable and automatic solution to detect ICMEs. The inconsistencies present in multiple catalogs serve as evidence for the difficulties that arise during this task.

We propose a pipeline that uses a well-known semantic segmentation algorithm for the segmentation of time series. It is trained, validated and tested on in situ data from Wind, reaching a True Skill Statistic of 0.64. Furthermore, reasonable performance could be achieved on datasets with fewer features and smaller training sets from Wind, STEREO-A and STEREO-B with True Skill Statistics of 0.56, 0.57 and 0.53, respectively.

In the future, the pipeline could be used for other time series event detection problems. Additionally, a conceivable extension would be the simultaneous detection of ICMEs and other structures by switching from binary segmentation to multiclass segmentation. In any case, combined with other forecasting methods, our tool can be expected to serve as a substantial contribution to space weather forecasting.

7 Data Availability Statement

The solar wind in situ data are available as Python numpy arrays at `https://doi.org/10.6084/m9.figshare.12058065.v7` and were originally downloaded from `https://stereo`

-ssc.nascom.nasa.gov (STEREO) and https://spdf.gsfc.nasa.gov/pub/data/wind/ (Wind).

The solar wind in situ data and catalog associated with Nguyen et al. (2019) can be found at https://github.com/gautiernguyen/Automatic-detection-of-ICMEs-at-1-AU-a-deep-learning-approach.

The HELIO4CAST ICMECAT catalog was used in the version 6 (updated on 2021 April 29). It is published on the data sharing platform figshare: https://doi.org/10.6084/m9.figshare.6356420.v6. The most up-to-date version can be found at https://helioforecast.space/icmecat.

The paper source code is available at https://github.com/ml-epn.

Acknowledgments
H.T.R., A.W., and U.V.A. thank Europlanet 2024 RI. Europlanet 2024 RI has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871149. C.M., T.A. and M.A.R. thank the Austrian Science Fund (FWF): P31521-N27, P31659-N27 and P34437.

References

Baumann, C., & McCloskey, A. E. (2021). Timing of the solar wind propagation delay between l1 and earth based on machine learning. *Journal of Space Weather and Space Climate*, *11*, 41. Retrieved from http://dx.doi.org/10.1051/swsc/2021026 doi: 10.1051/swsc/2021026

Burlaga, L., Sittler, E., Mariani, F., & Schwenn, R. (1981, August). Magnetic loop behind an interplanetary shock: Voyager, Helios, and IMP 8 observations. J. Geophys. Res., *86*(A8), 6673-6684. doi: 10.1029/JA086iA08p06673

Chen, J., Deng, H., Li, S., Li, W., Chen, H., Chen, Y., & Luo, B. (2022, March). RU-net: A Residual U-net for Automatic Interplanetary Coronal Mass Ejection Detection. ApJS, *259*(1), 8. doi: 10.3847/1538-4365/ac4587

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834-848. doi: 10.1109/TPAMI.2017.2699184

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking atrous convolution for semantic image segmentation.*

Chi, Y., Shen, C., Wang, Y., Xu, M., Ye, P., & Wang, S. (2016, October). Statistical Study of the Interplanetary Coronal Mass Ejections from 1995 to 2015. Sol. Phys., *291*(8), 2419-2439. doi: 10.1007/s11207-016-0971-5

Gosling, J. T., Pizzo, V., & Bame, S. J. (1973, January). Anomalously low proton temperatures in the solar wind following interplanetary shock waves—evidence for magnetic bottles? J. Geophys. Res., *78*(13), 2001. doi: 10.1029/JA078i013p02001

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *Lecture Notes in Computer Science*, 346–361. Retrieved from http://dx.doi.org/10.1007/978-3-319-10578-9_23 doi: 10.1007/978-3-319-10578-9_23

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.*

Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2019). *Squeeze-and-excitation networks.*

Jadon, S. (2020, oct). A survey of loss functions for semantic segmentation. In *2020 IEEE conference on computational intelligence in bioinformatics and compu-*

tational biology (CIBCB). IEEE. Retrieved from https://doi.org/10.1109 doi: 10.1109/cibcb48159.2020.9277638

Jha, D., Smedsrud, P. H., Riegler, M. A., Johansen, D., de Lange, T., Halvorsen, P., & Johansen, H. D. (2019). *Resunet++: An advanced architecture for medical image segmentation.*

Jian, L., Russell, C. T., Luhmann, J. G., & Skoug, R. M. (2006, December). Properties of Interplanetary Coronal Mass Ejections at One AU During 1995 2004. Sol. Phys., *239*(1-2), 393-436. doi: 10.1007/s11207-006-0133-2

Kilpua, E., Koskinen, H. E. J., & Pulkkinen, T. I. (2017, November). Coronal mass ejections and their sheath regions in interplanetary space. *Living Reviews in Solar Physics*, *14*(1), 5. doi: 10.1007/s41116-017-0009-6

Klein, L. W., & Burlaga, L. F. (1982, February). Interplanetary magnetic clouds at 1 AU. J. Geophys. Res., *87*(A2), 613-624. doi: 10.1029/JA087iA02p00613

Lepping, R., Acũna, M., Burlaga, L., Farrell, W., Slavin, J., Schatten, K., ... others (1995). The wind magnetic field investigation. *Space Science Reviews*, *71*(1), 207–229.

Lepping, R. P., Berdichevsky, D. B., Wu, C. C., Szabo, A., Narock, T., Mariani, F., ... Quivers, A. J. (2006, March). A summary of WIND magnetic clouds for years 1995-2003: model-fitted parameters, associated errors and classifications. *Annales Geophysicae*, *24*(1), 215-245. doi: 10.5194/angeo-24-215-2006

Li, H., & Guan, Y. (2021). Deepsleep convolutional neural network allows accurate and fast detection of sleep arousal. *Communications Biology*, *4*.

Lin, R. P., Anderson, K. A., Ashford, S., Carlson, C. W., Curtis, D. W., Ergun, R. E., ... Paschmann, G. (1995). A three-dimensional plasma and energetic particle investigation for the wind spacecraft. *Space Science Reviews*, *71*, 125-153.

Möstl, C., Weiss, A. J., Bailey, R. L., Reiss, M. A., Amerstorfer, T., Hinterreiter, J., ... Stansby, D. (2020, November). Prediction of the In Situ Coronal Mass Ejection Rate for Solar Cycle 25: Implications for Parker Solar Probe In Situ Observations. ApJ, *903*(2), 92. doi: 10.3847/1538-4357/abb9a1

Möstl, C., Weiss, A. J., Reiss, M. A., Amerstorfer, T., Bailey, R. L., Hinterreiter, J., ... Bale, S. D. (2022, January). Multipoint Interplanetary Coronal Mass Ejections Observed with Solar Orbiter, BepiColombo, Parker Solar Probe, Wind, and STEREO-A. ApJ, *924*(1), L6. doi: 10.3847/2041-8213/ac42d0

Nguyen, G., Aunai, N., Fontaine, D., Le Pennec, E., Van den Bossche, J., Jeandet, A., ... Regaldo-Saint Blancard, B. (2019, April). Automatic Detection of Interplanetary Coronal Mass Ejections from In Situ Data: A Deep Learning Approach. ApJ, *874*(2), 145. doi: 10.3847/1538-4357/ab0d24

Nieves-Chinchilla, T., Vourlidas, A., Raymond, J. C., Linton, M. G., Al-haddad, N., Savani, N. P., ... Hidalgo, M. A. (2018, February). Understanding the Internal Magnetic Field Configurations of ICMEs Using More than 20 Years of Wind Observations. Sol. Phys., *293*(2), 25. doi: 10.1007/s11207-018-1247-z

Ogilvie, K. W., Chornay, D. J., Fritzenreiter, R. J., Hunsaker, F., Keller, J., Lobell, J., ... Gergin, E. (1995, February). SWE, A Comprehensive Plasma Instrument for the Wind Spacecraft. Space Sci. Rev., *71*(1-4), 55-77. doi: 10.1007/BF00751326

Perslev, M., Jensen, M. H., Darkner, S., Jennum, P. J., & Igel, C. (2019). *U-time: A fully convolutional network for time series segmentation applied to sleep staging.*

Reiss, M. A., Möstl, C., Bailey, R. L., Rüdisser, H. T., Amerstorfer, U. V., Amerstorfer, T., ... Windisch, A. (2021, Dec). Machine learning for predicting the b z magnetic field component from upstream in situ observations of solar coronal mass ejections. *Space Weather*, *19*(12). Retrieved from http://dx.doi.org/10.1029/2021SW002859 doi: 10.1029/2021sw002859

Richardson, I. G., & Cane, H. V. (2010, June). Near-Earth Interplanetary Coronal

Mass Ejections During Solar Cycle 23 (1996 - 2009): Catalog and Summary of Properties. Sol. Phys., *264*(1), 189-237. doi: 10.1007/s11207-010-9568-6

Ronneberger, O., P.Fischer, & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention (miccai)* (Vol. 9351, pp. 234–241). Springer. Retrieved from `http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a` ((available on arXiv:1505.04597 [cs.CV]))

Rouillard, A. (2011). Relating white light and in situ observations of coronal mass ejections: A review. *Journal of Atmospheric and Solar-Terrestrial Physics*, *73*(10), 1201-1213. doi: 10.1016/j.jastp.2010.08.015

Rüdisser, H. T., & Amerstorfer, U. V. (2021, September). *Machine learning pipeline for automated detection of icmes.* `https://github.com/epn-ml/EPSC2021-ICME-workshop`. (Workshop held during the EPSC 2021)

Smith, L. N. (2017). *Cyclical learning rates for training neural networks.*

Topliff, C., Cohen, M., & Bristow, W. (2020). *Simultaneously forecasting global geomagnetic activity using recurrent networks.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need.*

Wilson, L., Brosius, A., Gopalswamy, N., Nieves-Chinchilla, T., Szabo, A., Hurley, K., ... TenBarge, J. (2021, 06). A quarter century of wind spacecraft discoveries. *Reviews of Geophysics*, *59*, e2020RG000714. doi: 10.1029/2020RG000714

Yang, J., Nguyen, M. N., San, P. P., Li, X., & Krishnaswamy, S. (2015). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Ijcai.*

Zhang, Z., Liu, Q., & Wang, Y. (2018). Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters*, *15*(5), 749-753. doi: 10.1109/LGRS.2018.2802944

Zurbuchen, T., & Richardson, I. (2006, 06). In-situ solar wind and magnetic field signatures of interplanetary coronal mass ejections. *Space Science Reviews*, *123*. doi: 10.1007/s11214-006-9010-4