MONTHLY NOTICES of the Royal Astronomical Society

An investigation of the causal relationship between sunspot groups and coronal mass ejections by determining source active regions

Journal:	*Monthly Notices of the Royal Astronomical Society*
Manuscript ID	MN-21-1582-MJ
Manuscript type:	Main Journal
Date Submitted by the Author:	30-Apr-2021
Complete List of Authors:	RAHEEM, Abd-ur; Canakkale Onsekiz Mart Universitesi, Graduate School of Natural and Applied Sciences Cavus, Huseyin; Canakkale Onsekiz Mart Universitesi, Department of Physics Coban, Gani Caglar; Canakkale Onsekiz Mart Universitesi, Graduate School of Natural and Applied Sciences Kinaci, Ahmet Cumhur; Canakkale Onsekiz Mart Universitesi, Department of Computer Engineering Wang, Haimin; New Jersey Institute of Technology, Centre for Solar-Terrestrial Research Wang, Jason T. L.; New Jersey Institute of Technology, Department of Computer Science
Keywords:	Sun: coronal mass ejections (CMEs) < The Sun, (Sun:) sunspots < The Sun, Sun: activity < The Sun

An investigation of the causal relationship between sunspot groups and coronal mass ejections by determining source active regions.

Abd-ur Raheem[1], Huseyin Cavus[2], Gani Caglar Coban[1], Ahmet Cumhur Kinaci[3], Haimin Wang[4] and Jason T. L. Wang[5]

[1]Canakkale Onsekiz Mart University Graduate School of Natural and Applied Sciences, 17110, Canakkale-TURKEY
e-mails: arawan@stu.comu.edu.tr, ganicaglarcoban@stu.comu.edu.tr
[2]Canakkale Onsekiz Mart University Department of Physics, 17110, Canakkale-TURKEY
e-mail: h_cavus@comu.edu.tr
[3]Canakkale Onsekiz Mart University Department of Computer Engineering, 17110, Canakkale-TURKEY
e-mail: cumhur.kinaci@comu.edu.tr
[4]New Jersey Institute of Technology, Center for Solar-Terrestrial Research, University Heights, Newark, NJ 07102-1982, USA
e-mail: haimin.wang@njit.edu
[5]New Jersey Institute of Technology, Department of Computer Science, University Heights, Newark, NJ 07102-1982, USA
e-mail: wangj@njit.edu

Abstract

Although the source active regions of some coronal mass ejections (CMEs) were identified in CME catalogues, vast majority of CMEs do not have an identified source active region. We propose a method that uses a filtration process and machine learning to identify the sunspot groups associated with a large fraction of CMEs and compare the physical parameters of these identified sunspot groups with properties of their corresponding CMEs to find mechanisms behind the initiation of CMEs. These CMEs were taken from the Coordinated Data Analysis workshops (CDAW) database hosted at NASA's website. The Helio-seismic and Magnetic Imager (HMI) Active Region Patches (HARPs) were taken from the Stanford University's JSOC database. The source active regions of the CMEs were identified by the help of a custom filtration procedure and then by training a Long Short-Term Memory Network (LSTM) to identify the patterns in the physical magnetic parameters derived from vector and line of sight magnetograms. The neural network simultaneously considers the time series data of these magnetic parameters at once and learns the patterns at the onset of CMEs. This neural network was then used to identify the source HARPs for the CMEs recorded from 2011 till 2020. The neural network was able to reliably identify source HARPs for 4895 CMEs out of 14604 listed in the CDAW database during the afore-mentioned period.

Keywords: Sun: activity, Sun: coronal mass ejections (CMEs), Sun: sunspots.

1. Introduction

Active regions are the patches of solar surface where there is an accumulation of strong magnetic field due to formation of sunspots. These patches are the source regions for solar activities, most prominently the solar flares and CMEs, as magnetic loops are intertwined and interacted to release free magnetic energy (Toriumi & Wang 2019).

The complex structure of the magnetic field within and among active regions allows plasma to get trapped around the magnetic loops that are formed among the closed magnetic field lines. These are observable and their height varies depending upon the complexity and strength of the source active regions. These magnetic loops can disintegrate, disconnect, rearrange, or undergo magnetic reconnection (Green 2016), a phenomenon where opposite magnetic field lines form new connections, to swirl plasma in the form of clouds into the interplanetary environment called CMEs. The CMEs can accelerate or decelerate near solar surface in the interplanetary environment depending on the conditions of their origin and their interaction with the solar wind. Initially fast CMEs are observed to decelerate in their propagation afterwards (Manoharan 2006). CMEs are sources of various space weather effects in the near-earth environment, such as geomagnetic storms, . The propagation of CMEs can keep on going past 1 AU as was observed by Ulysses mission (Richardson 2014) and the effects can be seen around Mars (von Forstner et al. 2018).

Active Regions on the solar surface, are directly linked to magnetic activity on the solar surface and thus to flares, CMEs, interplanetary shock waves and other activities. The evolution in the magnetic field of these regions directly effects the initiation of these solar phenomena. Understanding the evolution and structure of magnetic is the key to the prediction of flares and CMEs. Various studies have been carried out for solar flare and coronal mass ejection predictions using machine learning (see, Yan, Qu & Kong 2011; Bobra & Ilonidis 2016). These techniques include the prediction of the CMEs related to solar flares and solar emission particles (e.g., Chandra et al. 2015; Liu et al. 2019; Liu et al. 2020).

This study attempts to define a method to relate the sunspot groups with CMEs so that magnetic parameters of such sunspot groups can be compared to the respective CMEs to investigate the mechanisms for the onset of CMEs. Machine learning, especially neural networks, has been employed before for the studies of, but not limited to, interplanetary environment, space weather forecast, and interplanetary shocks (Cavus et al. 2020). CMEs are in fact the primary sources for geomagnetic storms and interplanetary shocks (Gosling 1993; Hudson & Ryan 1995). In existing CME catalogues, only a small fraction of them have identified source regions. Our goal is to identify as many as possible the source ARs of CMEs. Machine learning based on the relationship between magnetic parameters in existing identified ARs and associated CMEs is used in this study.

2. Material and Method

2.1. Data mining and preparing the datasets:

The CME information is taken from the CDAW database that uses data from the Large Angle and Spectrometric Coronagraph (LASCO) onboard NASA's Solar and Heliospheric Observatory (SOHO) spacecraft (URL-1). The AR magnetic properties are extracted from data

of Stanford University's Joint Science Operations Centre (JSOC). JSOC maintains databases of Space-weather HMI Active Region Patches (SHARP). These HMI Active Region Patches will be referred to as HARPs from this point onwards. A variety of magnetic field parameters for these HARPs are derived from the vector magnetogram data (URL-2). These quantities are given below in Table 1.

Parameter	Formula	Description (unit)
MEANPOT	$\overline{\rho} \propto \frac{1}{N}\Sigma(B^{Obs} - B^{Pot})^2$	Mean photospheric excess magnetic energy density (erg/cm^3)
TOTPOT	$\rho_{tot} \propto \Sigma(B^{Obs} - B^{Pot})^2 dA$	Total photospheric magnetic free energy density (erg/cm)
USFLUX	$\Phi = \Sigma\lvert B_z\rvert dA$	Total unsigned flux (Mx)
MEANGAM	$\overline{\gamma} \propto \frac{1}{N}\Sigma \arctan(\frac{B_h}{B_z})$	Mean inclination angle (Deg)
MEANGBT	$\lvert\overline{\nabla B_{tot}}\rvert = \frac{1}{N}\Sigma\sqrt{\left(\frac{\partial B}{\partial x}\right)^2 + \left(\frac{\partial B}{\partial y}\right)^2}$	Mean gradient of total field (G/Mm)
MEANJZD	$\overline{J_z} \propto \frac{1}{N}\Sigma(\frac{\partial B_y}{\partial x} - \frac{\partial B_x}{\partial y})$	Mean vertical current density (mA/m^2)
TOTUSJZ	$J_{Ztotal} = \Sigma\lvert J_z\rvert dA$	Total unsigned vertical current (A)
MEANGBH	$\lvert\overline{\nabla B_h}\rvert = \frac{1}{N}\Sigma\sqrt{\left(\frac{\partial B_h}{\partial x}\right)^2 + \left(\frac{\partial B_h}{\partial y}\right)^2}$	Mean gradient of horizontal field (G/Mm)
MEANGBZ	$\lvert\overline{\nabla B_z}\rvert = \frac{1}{N}\Sigma\sqrt{\left(\frac{\partial B_z}{\partial x}\right)^2 + \left(\frac{\partial B_z}{\partial y}\right)^2}$	Mean value of the vertical field gradient, in G/Mm
MEANALP	$\alpha_{total} \propto \frac{\Sigma J_z B_z}{\Sigma B_z^2}$	Mean twist parameter, α (Deg)
MEANJZH	$\overline{J_z} \propto \frac{1}{N}\Sigma B_z J_z$	Mean current helicity (G^2/m)
TOTUSJH	$H_{Ctotal} \propto \Sigma\lvert B_z \cdot J_z\rvert$	Total unsigned current helicity (G^2/m)
ABSNJZH	$H_{C\,abs} \propto \lvert\Sigma B_z \cdot J_z\rvert$	Absolute value of the net current helicity (G^2/m)
SAVNCPP	$J_{Zsum} \propto \lvert\Sigma^{B_z{}^+} J_z dA\rvert + \lvert\Sigma^{B_z{}^-} J_z dA\rvert$	Sum of the absolute value of the net current per polarity (A)
MEANSHR	$\overline{\Gamma} = \frac{1}{N}\Sigma\arccos\left(\frac{B^{\mathrm{Obs}} \cdot B^{\mathrm{Pot}}}{\lvert B^{\mathrm{Obs}}\rvert\lvert B^{\mathrm{pot}}\rvert}\right)$	Mean shear angle (Deg)
SHRGT45	Area with shear greater than 45°/total area	Area with shear angle greater than 45 degrees (percent of total area)
R_VALUE	$\Phi = \Sigma\lvert B_{LOS}\rvert dA$ within R mask	Flux along gradient-weighted neutral-line length (Mx)

Table 1. The HMI magnetic parameters obtained from JSOC used as inputs in this study.

The parameters of the photospheric magnetic field are correlated with the solar activity (Falconer et al. 2002; Leka and Barnes 2003a and 2003b; Schrijver 2007). The SDO is in an inclined geosynchronous circular orbit (IGSO) at approximately 35,756 km above Earth whereas SOHO is at Lagrange point L1 that is approximately 1.5 million km away from Earth.

The CME database (CDAW database hereafter) records all the events that are declared as CMEs. The database has parameters including the time, the central positional angle, width, linear speed, 2nd order initial speed, 2nd order final speed, 2nd order speed at 20R, acceleration of the CME, the mass, kinetic energy, and the mean positional angle of each recorded CME. Moreover, there is a comment section for each CME which indicates the quality of the event

i.e., good event or poor event based on the number of spatial points available for each observation and to measure the parameters because in some cases the number of data points available is small and the quality of the measurement suffers. The parameters e.g. energy, are derived from the CME speed and estimated mass (Gopalswamy et al. 2009). In some cases, there is no measurement available for parameters including CME speed or mass. This catalogue combines results from the different chronograph instruments available in LASCO, namely, LASCO C1, LASCO C2 and LASCO C3 (URL-1).

A list of previously known active regions which have initiated events including CMEs or solar flares is being hosted at The Space Weather Database of Notifications, Knowledge, Information (DONKI)(URL-3) of The Community Coordinated Modelling Centre (CCMC). This provides us an opportunity to train a neural network to learn the patterns between the data of these events and then the trained network can be used to find all the related events in the previous years which have not been currently labelled/related. Of all the cases in the DONKI database 156 of them are CME events with known source regions. The above-mentioned magnetic parameters of 120 of these events could be traced and fetched from the SHARP database. The data from the SHARP database was obtained through the JSOC's API (also available through SunPy) (Sunpy community et al. 2015). Datasets were formed that contain all mentioned magnetic parameters from the source regions of these CMEs along with their respective CMEs. At this stage, an analysis could be performed on the data by a neural network. This is achieved by training models including Random Forest, SVM, kNN, Decision tree and LSTM on the data that has been previously labelled. The events were taken from DONKI and the data for these events were taken from the SHARP (for active regions) database and CDAW (for CMEs) database.

The data of input parameters includes different lengths of time based on the dataset used i.e., 2 hours, 4 hours, 6 hours, 8 hours, 10 hours, and 12 hours. There are cases where the used data for HARPs have additional CMEs during the included time-period that are not listed in the DONKI database but are listed in the CDAW database. This implies presence of some other CMEs at the same time and in the same positional angle vicinity as the used HARP so data for such HARPs were omitted from the initial dataset used for training as these HARPs may be associated with the mentioned additional CMEs in the used dataset. Such data has points in the dataset which cannot be definitively labelled as either CME or NOT CME. Thus, the number of the HARPs used in the training procedure for the LSTM was reduced depending on the length of the dataset as the probability of the presence of such points increases with the length of data used for each HARP. Table 2 below shows the final number of HARPs left after the omission of such HARPs. The length of each dataset is given in an array structure (e.g. (780, 5, 17) where 780 represents the number of total data points in the different timeseries given to the model, 5 represents the timestep meaning the 5 points are given at the same time to the model while training and then window slides one step in the forward direction, 17 represents the number of features or the number of different timeseries given representing 17 different but co-dependent features). This structure in fact helps show the length and configuration of the datasets.

Dataset	Length of dataset	Number of HARPs used	Number of data samples labelled 'CME'	Number of data samples labelled 'NOTCME'
2-hour dataset	(780,5,17)	78	78	702
4-hour dataset	(1200,5,17)	60	59	1141
6-hour dataset	(1320,5,17)	44	44	1276
8-hour dataset	(1240,5,17)	31	31	1209
10-hour dataset	(1150,5,17)	23	23	1127
12-hour dataset	(1200,5,17)	20	20	1180

Table 2. An overview of the datasets used for the machine leaning models.

The time series data was normalized within each active region using [–1 to 1] scaling. This is done so the network can catch the patterns and not learn the specific values of the features. If not normalized the weights and bias of the network are severely affected due to the difference in the scale of the features between different active regions. By normalizing data this way, the features having significant differences in the minimum and maximum values are treated in an equivalent manner. This enables the creation of models that can generalize the problem otherwise the models are not able to maintain the weights within a mathematically usable limit. The weights quickly explode while the model tries to fit the data. Some data has missing time steps where data is not calculated and therefore there are data gaps in the time series. These gaps were filled with averages of data before and after each gap. The amount of such filler data is less than 3%. Moreover, in some cases a padding was included in the time series data of some active regions to maintain the dimensions of the time series in preparation for the LSTM network. In order to include cases where there is no CME present, time periods where there were no CMEs were selected carefully verifying from the SHARP and CDAW data.

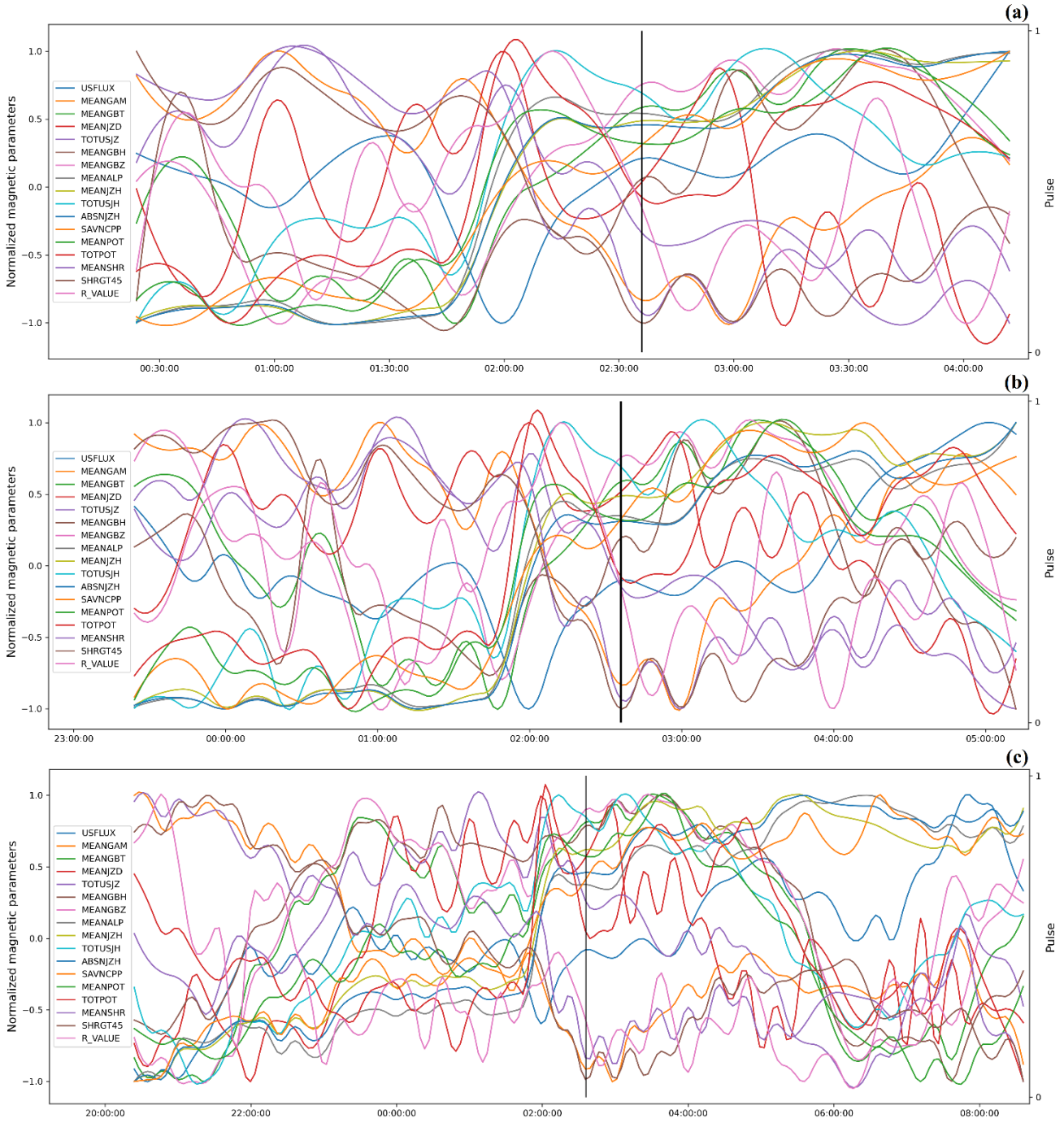

Figure 1. A representation of the pulse generated for the HARPs in the datasets. (a) shows the first HARP from the 6-hour dataset and (b) shows the first HARP from the 12-hour dataset. (c) shows the first HARP from the 12-hour dataset along with its pulse. The y-axis on left shows the magnetic parameters of the depicted HARPs and y-axis on the right shows the generated pulse for the HARP. Pulse is binary in the dataset and shown here as a bar in black. The active region 11158 appeared on 15th February 2011 is displayed in the figure where the number is assigned by National Oceanic and Atmospheric Administration (NOAA). The CME for the pulse shown here appeared on 15th February 2011.

For the labelling of the datasets, a parameter called pulse is formed which reflects the presence of a CME or the absence in binary. The pulse takes the value of 1 when CME is

present at the corresponding time else it takes the value of 0. Moreover, the cadence of the SHARP series used for active regions is 12 minutes(hmi.sharp_cea_720s). Some CMEs fall in between two consecutive datapoints of the active region data. To overcome this CME time have been shifted to round to multiples of 12 in order to calculate pulse for these cases. Fig. 1 shows first HARPs from the 4-hour, 6-hour and the 12-hour dataset as an example. The magnetic parameters of the HARP and the generated pulse is shown for those magnetic parameters in Fig. 1. The datasets were created so that usually the CME falls in the middle. So, the 2-hour dataset would be ±1 hour with respect to the CME, ±2 hours for 4-hour dataset and so on. There are some HARPs in datasets where there is no CME and there are some samples where there are 2 CMEs. The number of datapoints increase in the datasets (from 2 hours towards 12 hours) because of an increase in the length of elapsed time as shown in Fig. 1.

When feeding the networks, the data samples are created by sliding a window forward. This window can be thought of as fixed number of consecutive data samples from the datasets. The window (e.g., will span 5 data points that corresponds to 1 hour) slides skipping one data sample at a time i.e., 5 data samples are taken from the start at first. The target label (value of parameter 'pulse') is selected based upon the last data point in that set of data samples. So, in fact it answers the question: 'Does the change in this set of data points result in a CME?' The window starts from the beginning of a dataset till the end. The datasets in this manner are transformed into timeseries data suitable for machine learning. For predictions, the pulse for each data sample is predicted and this process in reverse is applied to get the predicted source region for that CME.

2.2. Using LSTM to learn the patterns in the magnetic parameters of HARPs.

The objective was to find the patterns in the magnetic features of the active regions to identify the initiation of CMEs and thus find the source regions for CMEs identified from the years between 2011 and 2020. The magnetic parameters mentioned above make a set of co-dependant time series, which in return becomes near impossible to analyse without the employment of any machine learning algorithm. The problem here is to identify a set of changes in these parameters simultaneously with respect to the initiation of CME and then use the acquired information to identify source regions for other CMEs. However, the magnitude of the effect caused by individual changes within the set of identified changes is essential to correctly predict the source regions and thus machine learning is employed in this matter frequently (see, Bobra & Ilonides 2016; Inceoglu et al. 2018). Hence, this study tries to use a machine learning technique, called Long Short-Term Memory neural network LSTM, to produce a model that can successfully learn these patterns and then the model can be employed to identify source regions for a large number of CMEs.

Five different machine learning algorithms were employed in this study to identify the best algorithm for the problem at hand. These include Decision Tree, Support Vector Machines (SVM), Random Forest (RF), k-nearest neighbour (kNN) and LSTM. The datasets were divided to form training and test datasets. 75% of the data were used for training and 25% of the data were used as test dataset. These datasets do not include common data and were used for evaluation of the models. Due to the presence of an imbalance in the frequency of the classes i.e., CME class and NOTCME class in the used datasets (Table 2), it is impossible to train any

model as the models regard the CME class as noise and do not learn to distinguish between the classes. The ratio of CME class to NOTCME class is 78:702 (11%) for the 2-hour dataset, 59:1141 (4.9%) for 4-hour and so on. To overcome this problem, in the data, weights were applied to the two classes present in the data depending on their frequency of occurrence within the datasets. This implementation rewards and penalizes the models differently for different classes and forces the models to not regard the smaller class as noise depending on the frequency of the class in the data. For Decision Tree the minimum number of instances in leaves, the minimum split subset, and the maximal tree depth were set to 95, 11 and 4, respectively. The minimum number of neighbours was set to 17 as this generally gave the optimal results. The weight metric was determined to be Manhattan distance in the used kNN models. The number of attributes considered at each split, depth of individual trees and the smallest split size for Random Forest were set to 5, 3 and 5, respectively. The cost was set at 1 for SVM and iteration limit was determined to be 100.

Results produced by the best models based on these techniques are given in Table 3. The results show that the complexity of the problem is quite high for Decision Tree, SVM, RF and kNN as these models were not able to produce reliable and acceptable results. The results shown in Tables 3 and 6 depict a large contrast between these techniques and the LSTM network.

	Model	Precision	Recall	F1 score	Accuracy
2-hour data	**Decision Tree**	0.59	0.59	0.58	0.63
	SVM	0.50	0.47	0.43	0.47
	kNN	0.63	0.61	0.59	0.61
	Random Forest	0.64	0.63	0.63	0.63
4-hour data	**Decision Tree**	0.63	0.62	0.62	0.62
	SVM	0.51	0.50	0.48	0.51
	kNN	0.59	0.56	0.58	0.59
	Random Forest	0.58	0.56	0.54	0.56
6-hour data	**Decision Tree**	0.61	0.60	0.60	0.60
	SVM	0.57	0.54	0.51	0.54
	kNN	0.44	0.45	0.43	0.45
	Random Forest	0.65	0.65	0.64	0.65
8-hour data	**Decision Tree**	0.33	0.33	0.32	0.37
	SVM	0.44	0.45	0.38	0.45
	kNN	0.52	0.50	0.47	0.50
	Random Forest	0.63	0.62	0.61	0.62
10-hour data	**Decision Tree**	0.43	0.44	0.39	0.44
	SVM	0.21	0.44	0.29	0.44
	kNN	0.45	0.45	0.40	0.45
	Random Forest	0.47	0.46	0.41	0.46
12-hour data	**Decision Tree**	0.51	0.51	0.50	0.51
	SVM	0.51	0.51	0.51	0.51
	kNN	0.48	0.48	0.44	0.48
	Random Forest	0.58	0.57	0.55	0.57

Table 3. Performance of different machine learning algorithms used on test data for different datasets used in this study.

These results highlight the fact that the number of features for each sample is too large for these used machine learning algorithms and thus a more sophisticated technique i.e., LSTM was employed for this study. These algorithms form base models, and better results were obtained by LSTM model described in detail later in this section. The performance of the LSTM model used is given at the end of this section. Table 4 shows the confusion matrix for the respective results given above where the prediction of the models is shown vertically and the actual data of those predictions horizontally. Results for all the test datasets are included in the table which highlights the performance of these model based on target classes i.e., CME and NOTCME which in this case represent the presence and absence of CMEs, respectively.

Algorithm				**Predicted**	
				NOTCME	CME
2-hour data	Decision Tree	NOTCME	Actual	58.3%	41.2%
		CME		41.7%	58.8%
	SVM	NOTCME		47.4%	52.6%
		CME		52.6%	47.4%
	kNN	NOTCME		69.3%	41.6%
		CME		30.7%	58.4%
	Random Forest	NOTCME		59.3%	30.2%
		CME		40.7%	69.8%
4-hour data	Decision Tree	NOTCME		61.0%	35.0%
		CME		39.0%	65.0%
	SVM	NOTCME		50.1%	47.7%
		CME		49.9%	52.3%
	kNN	NOTCME		57.7%	40.7%
		CME		42.3%	59.3%
	Random Forest	NOTCME		54.0%	38.7%
		CME		46.0%	61.3%
6-hour data	Decision Tree	NOTCME		59.4%	37.2%
		CME		40.6%	62.8%
	SVM	NOTCME		52.8%	38.7%
		CME		47.2%	61.3%
	kNN	NOTCME		46.4%	56.7%
		CME		53.6%	43.3%
	Random Forest	NOTCME		62.2%	30.4%
		CME		37.8%	69.6%
8-hour data	Decision Tree	NOTCME		40.5%	72.9%
		CME		59.5%	27.1%
	SVM	NOTCME		46.6%	57.6%
		CME		53.4%	42.4%
	kNN	NOTCME		49.2%	45.1%
		CME		50.8%	54.9%
	Random Forest	NOTCME		58.3%	31.9%
		CME		41.7%	68.1%
10-hour data	Decision Tree	NOTCME		44.9%	58.1%
		CME		55.1%	41.9%
	SVM	NOTCME		45.7%	100%
		CME		54.3%	0.0%
	kNN	NOTCME		45.9%	55.0%
		CME		54.1%	45.0%
	Random Forest	NOTCME		46.6%	52.4%

		CME		53.4%	47.6%
12-hour data	Decision Tree	NOTCME		51.0%	48.2%
		CME		49.0%	51.8%
	SVM	NOTCME		51.4%	47.6%
		CME		48.6%	52.4%
	kNN	NOTCME		49.1%	52.8%
		CME		50.9%	47.2%
	Random Forest	NOTCME		55.3%	38.9%
		CME		44.7%	61.1%

Table 4. Confusion matrix for the produced models of Decision Tree, SVM, kNN and RF.

There are some active regions for which there is no data available in SHARP. These were omitted from the analysis. The datasets prepared with various lengths were all trained in order to determine the best dataset on which a model could be trained. The objective here was to maximize the performance of the model.

LSTM networks are proven to be capable and very efficient in learning patterns in time series data and the classification of time series data. (Karim et al. 2019). The output layer of the network was categorized so that there exist two classes one for the presence of CME and the other for the absence of the CME. For this, one-hot-encoding was performed to categorize the presence of CME in the input data and vice versa. For LSTM, 70% of the data were used as training data and 15% of the data was used as validation during the training of the LSTM model to tune the hyperparameters. The remaining 15% of the dataset was used as test dataset used for testing the model after the training. LSTM networks were programmed in python using Tensorflow and Keras. Table 5 shows the lengths of the datasets for the 4-hour dataset as an example. The format is same as the format used to describe the features of datasets used discussed in Section 2.1 and Table 2.

Type	Size
Total Dataset	(1200,5,17)
Total training dataset	(837,5,17)
Total validation dataset	(179,5,17)
Total test dataset	(179,5,17)

Table 5. An example for the datasets used for the LSTM model. This example depicts the division of 4-hour dataset.

LSTM networks have a window (also known as the lookback of an epoch) that is moved through the length of a time series data while training for output data/labels to learn the patterns. The window/lookback was tuned according to the performance of the data and the performance were discovered to be at one hour. The hyperparameters of the models were tweaked for different datasets used in this study to optimize for performance. The LSTM network for the 10-hour dataset has 150 neurons in the LSTM layer, the input and output layers are designed according to the input and output of the network while the 4-hour dataset has 50 neurons. These parameters change for different datasets as they perform differently with different hyperparameters with some having different number of dense layers. The activation function used in the LSTM layer is ‘tanh’. The categorical cross-entropy with logits in the layer was used as the loss function as the data used is categorical. The Adaptive Moment Estimation

'Adam' (Kingma & Ba 2015) was used as optimizer, a method for stochastic gradient descent, which produced the most reliable and consistent training results. Class weights were added to the CME and NOTCME classes as there is a severe imbalance (consult Table 2). The weight was calculated based on the frequency of classes in their respective datasets. An early stopping criterion was added over the evaluation function to save time in the training process and to avoid overfitting.

A dropout amount was set to 0.2 in the training process of the LSTM networks to avoid overfitting. Fig. 2 shows the loss of the model based on the number of epochs; the loss continuously reduces until the early stopping criterion kicks in due to absence of meaningful further improvement in the quality of predictions during that training session. A patience setting of 5 for the early stopping criterion was determined to be optimal for this study. Table 6 shows the performance of LSTM models during training on training and validation data while Table 7 shows the results of the LSTM models for different lengths of datasets on test datasets. Since, the CMEs were shifted forward to multiples of 12 in minutes where necessary in order to compensate for the cadence (12 minutes) of the SHARP series used for active regions, CME predictions with an error no more than 12 minutes were considered correct. The results for the model used later in the study to identify source active regions are given in bold. Note that the results here are all higher than the results obtained from other algorithms used given in Table 3. It was determined based on these results that the 4-hour dataset produces the best results based on accuracy, precision and f1 score. This dataset was used for the prediction of the source regions for CMEs between 2011 and 2020 discussed in the next section. The importance of the precision metric for this study is discussed further in the results section.

Metric		**Size of the training data**					
		2-hour	**4-hour**	**6-hour**	**8-hour**	**10-hour**	**12-hour**
Accuracy	training	0.96	**0.97**	0.95	0.82	0.91	0.87
	Validation	0.87	**0.90**	0.85	0.72	0.94	0.85
	All	0.93	**0.94**	0.92	0.80	0.89	0.87

Table 6. Performance of different LSTM models on different datasets used in this study during training.

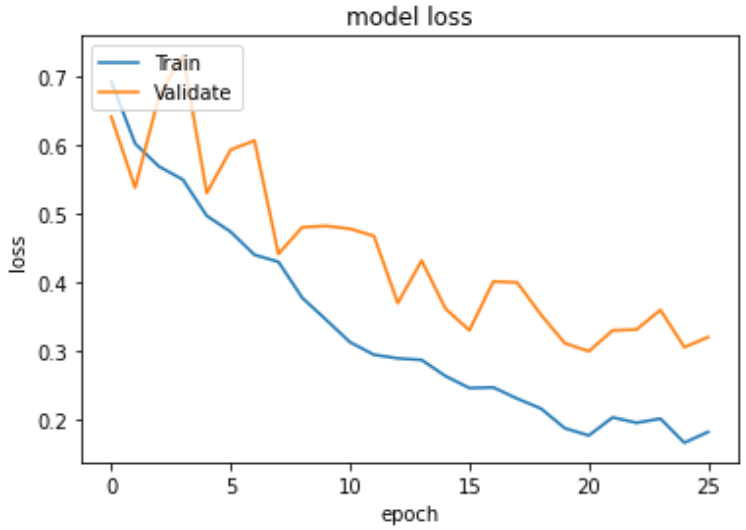

Figure 2. The loss of the LSTM model vs the epoch during training.

Result Metric	Size of the training data					
	2-hour	4-hour	6-hour	8-hour	10-hour	12-hour
Accuracy	0.80	**0.86**	0.87	0.77	0.71	0.81
Recall	0.79	**0.77**	0.86	0.81	0.86	0.83
Precision	0.65	**0.81**	0.52	0.38	0.32	0.25
F1 score	0.72	**0.79**	0.65	0.54	0.46	0.39

Table 7. Results for different LSTM models based on different datasets used in this study. Boldface is used for the model used later for the identification of source active regions.

The accuracy, precision and recall of the selected model is 86%, 81%, 77% respectively i.e., the model trained on 4-hour data. This model is considered well trained to be used for the task of determining the source ARs for each CME event from 2011 till 2020. A time-based representation of the pulse is shown in Fig. 3. Each line in the graphs in Fig. 3 represents a CME in the dataset hence the y-axis shows 1 or 0 depending on the presence or absence of CME. The dispersion of datasets on the x-axis shows the shuffling of used data for the training, validation, and test. A detail is given in the legend. Fig. 3 depicts the model's ability to generalize the problem and an absence of overfitting the test dataset is also depicted which was not fed to the models during training. This LSTM model was used later for further optimization and the identification of source regions for CMEs from 2011 till 2020 mentioned in the next section.

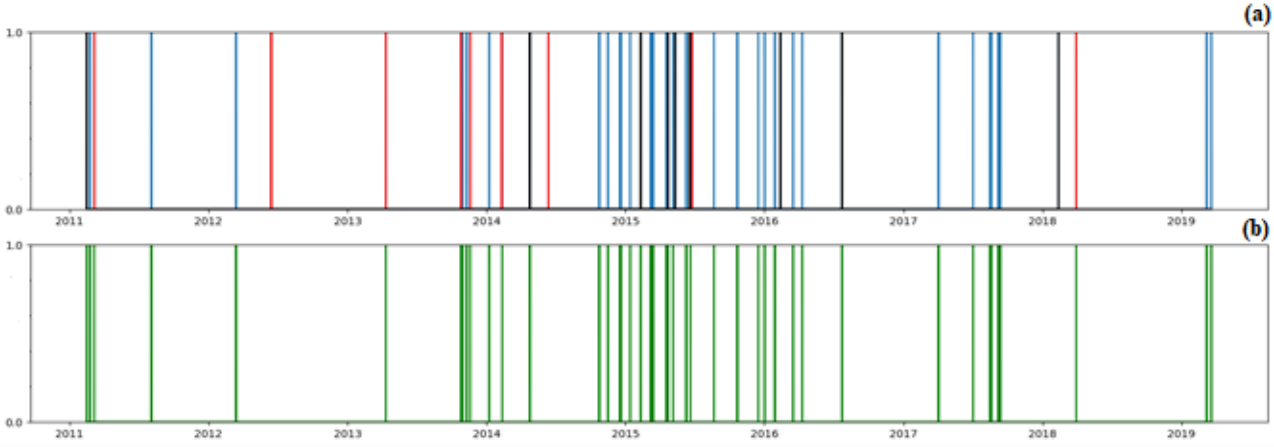

Figure 3. The performance of the selected model during and after training on different datasets. Here training data is represented in blue, the validation data in black and test data is represented in red. The prediction of the LSTM is depicted in green. (a) shows the actual dataset and (b) shows the predictions made by the LSTM.

2.3. AR-Pool for each CME and predictions using the LSTM network.

Before making predictions on the data between 2011-2020, the ARs were passed through the following filter for each CME. The filter can be described as having a predefined criterion for the HARPs with respect to a particular selected CME. We divide the filtration criteria in the following steps.

For a selected CME:

1) All ARs that are present at the same time as the CME are selected.
2) Within those ARs those within ± 90° of positional angle as the CME are selected.
3) Within the selected ARs in stage 2, only those that spend at least 4 hours in the area as determined in the stage 2 of the filter are selected.

To achieve the above-mentioned filtration, process the mean positional angle (MPA) for each HARP was calculated from the average values of the latitude and longitude of the HARP at a particular instance. Some morphologic calculations were performed and the whole process was automized for the whole SHARP database. The third stage in the filter is designed in mentioned manner so that the data is suitable for LSTM network to perform predictions. Also, the ARs which happen to be at the edges of the determined area for each CME in stage 2 could be ignored this way as they continue to displace to an area that cannot be correlated to the initiation of that CME. In this way several ARs responsible for the initiation of each CME from 2011 to 2020 were selected. The number of these ARs vary from 3 to 13 depending upon the CME and the solar cycle. For some CMEs in 2014 during the solar maximum the maximum number of responsible ARs determined is 13 whereas for some CMEs in 2019 during the solar minimum the maximum number of responsible ARs selected for that CME is 3. Since, more than one AR could be associated with each CME after this step an AR-pool was formed.

After this process, the data of these ARs from the AR-pool for each CME were fed to the model and predictions were made to narrow down the selected ARs from the AR-pool for each CME. The normalization and data modification steps performed on the data are the same as those for the data used during the formation of the LSTM model.

3. Results and Discussion

Table 8 shows results based on the filtration process and Table 9 shows results obtained after the data obtained from the filtration process was subjected to the LSTM prediction process.

The total number of CMEs found the CDAW database used in this study from 2011 till the 8th month of 2020 is 14604. Of these 12451 were associated with Active Regions based on the filtration process discussed in Section 2.3 (Table 8). This formed an AR-pool for their corresponding CMEs. These cases were then subjected to the LSTM model prediction process to narrow down the source AR-pool for each CME. The 'total predictions made' column shows the total predictions made by the LSTM model including errors and cases where more than one HARP was selected by the model from the AR pool. In 4895 cases a single AR responsible for the initiation of the CME was determined are given in Table 9. These are the cases where the neural network only chose one AR from the AR-pool of the corresponding CME. A database of these CMEs along with their respective AR is consequently produced in this study. The produced database and data used in this study is openly available (Raheem et al. 2021).

Year	**CMEs**	
	Total CMEs listed in CDAW	For which an AR-pool could be formed
2011	1990	1817
2012	2177	1955
2013	2338	2183

2014	2477	2323
2015	2057	1894
2016	1392	1251
2017	785	565
2018	475	193
2019	548	144
2020	365	126

Table 8. Statistics after the filtration process also discussed in Section 2.3.

Year	Total Predictions made	Cases where the LSTM model selected one HARP from the AR-pool
2011	873	616
2012	1124	700
2013	1352	781
2014	1817	924
2015	1192	715
2016	682	468
2017	565	325
2018	293	151
2019	214	107
2020	186	108

Table 9. Statistics after data of filtered ARs were subjected to the LSTM prediction process (URL-4).

The relationship between the magnetic parameters of the identified source regions with their respective CMEs is analysed from the database produced as an outcome of this study. The change in some current related parameters (i.e., MEANJZD, TOTUSJZ and SAVNCPP) of the source regions with respect to the linear speed, acceleration, mass, and kinetic energy of the initiated CME, since these are motion related, is given in Fig. 4. The data was normalized between 0 and 1 to overcome the obvious large differences between the ranges of these parameters. Acceleration of CMEs and MEANJZD of the source regions were normalized between -1 and 1 due to the presence of negative values in original data. The results of these analyses can be summarized as the following:

1. The minimum and maximum values of the mean vertical current density (MEANJZD) of HARPs in the SHARP database are given as -4.39 and 7.69 mA/m^2, respectively. The mean vertical current density of the identified source regions for their respective CMEs is between -1.39 mA/m^2 and 2.86 mA/m^2. This implies that only source regions within this range seem to be playing a role in the initiation of CMEs. Similarly, the value of the total unsigned vertical current (TOTUSJZ) of the identified source regions is around $6.5e^{+13}$ A whereas, the average value of the parameter is around $7.6e^{+12}$ A in the SHARP database. The sum of the absolute value of the net current per polarity (SAVNCPP) of the identified source regions for the CMEs is generally around $3e^{+13}$A (consult Fig. 4 and URL-4).
2. The CMEs initiated by these source regions have linear speeds below 1058.54 km/s, masses below $1.33e^{+16}$ g and kinetic energies normally around $6.20e^{+31}$ erg (consult Fig. 4(a) through Fig. 4(l)).

3. Fig. 4 highlights the fact that the kinetic energy of the CMEs and the total unsigned vertical current of their source regions are inversely proportional to each other and have an asymptotic relation with each other. The same can be said for the sum of the absolute value of the net current per polarity of the source regions and the kinetic energies of CMEs initiated by them. This can be observed in Fig. 4(k) and Fig. 4(l).
4. Generally, the mass of CME is inversely proportional to the sum of the absolute value of the net current per polarity and the total unsigned vertical current of its source region (consult Fig. 4(h) and Fig. 4(i)).
5. Accelerating CMEs usually originate from source regions having a negative mean vertical current value according to results obtained in this study (see, Fig. 4(d)).
6. The average momentum of the CMEs for which a source region could be identified is $7.88e^{+17}$ gkm/s. 84.42% of the CMEs in the database are below this average value.

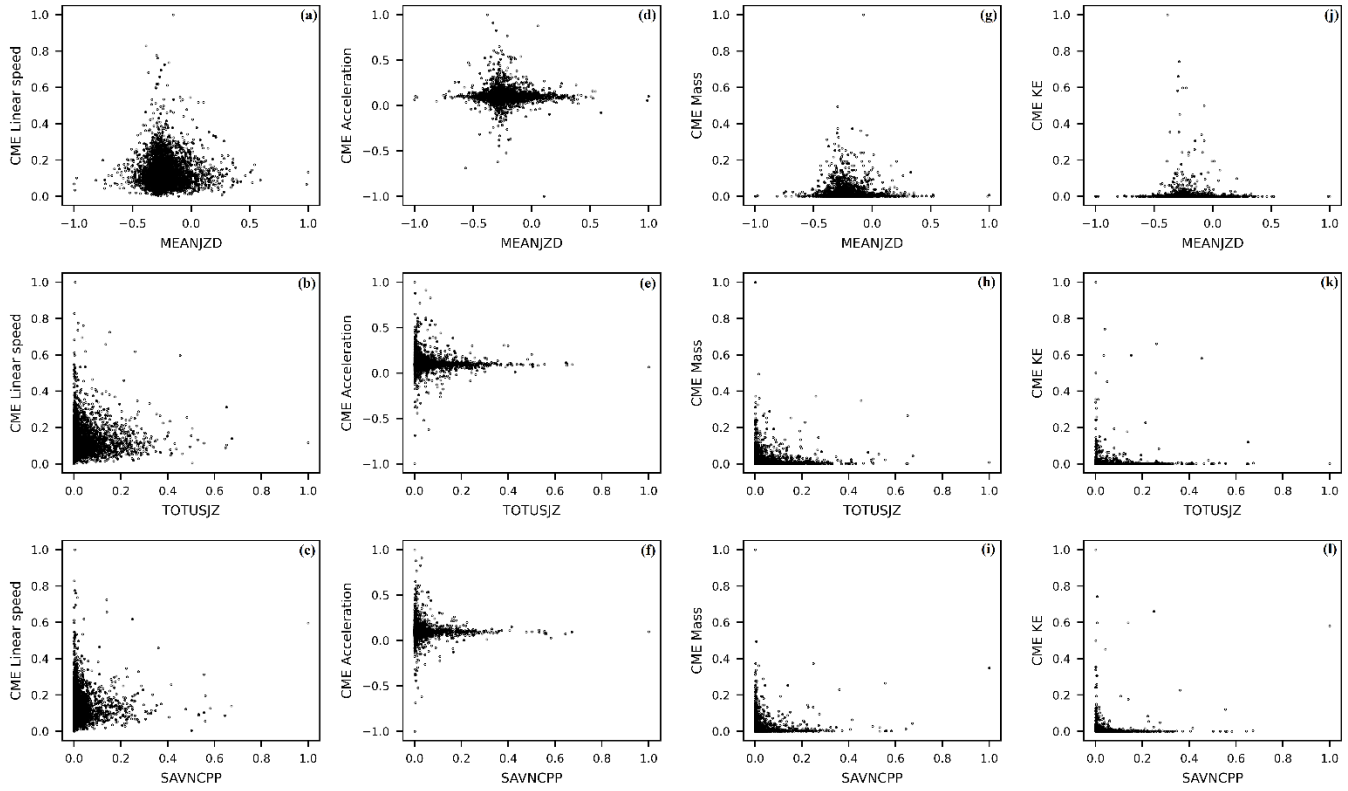

Figure 4. The relationship between linear speed, acceleration, mass, and KE of CMEs with MEANJZD, TOTUSJZ and SAVNCPP of their source regions.

These results are based on the predictions of the LSTM network which was trained on a very limited data present in the DONKI database. The uncertainty of the database is not quantified. There also exists noise in the data obtained from the DONKI database and the parameters then obtained from SHARP database. However, the model behaves in a desired manner as we tried to keep the precision value as high as possible without signs of overfitting. The recall metric was of low priority during the training as the correctness of the predictions was of utmost importance. Determination of high number of correct labels of CMEs rather than a high frequency of labels determined were aimed during the design of the models. This was set so that when the predictions were made on the data from 2011 to 2020 the confidence level

of prediction could be kept high. Therefore, the model makes less predictions as compared to the total number of CMEs present during that timeframe. These prediction numbers and quality can be enhanced if more data is fed to the model which at this stage is not possible. This study is unique as it is the first attempt to our knowledge at creation of an automated large scale identification method of the source regions for CMEs.

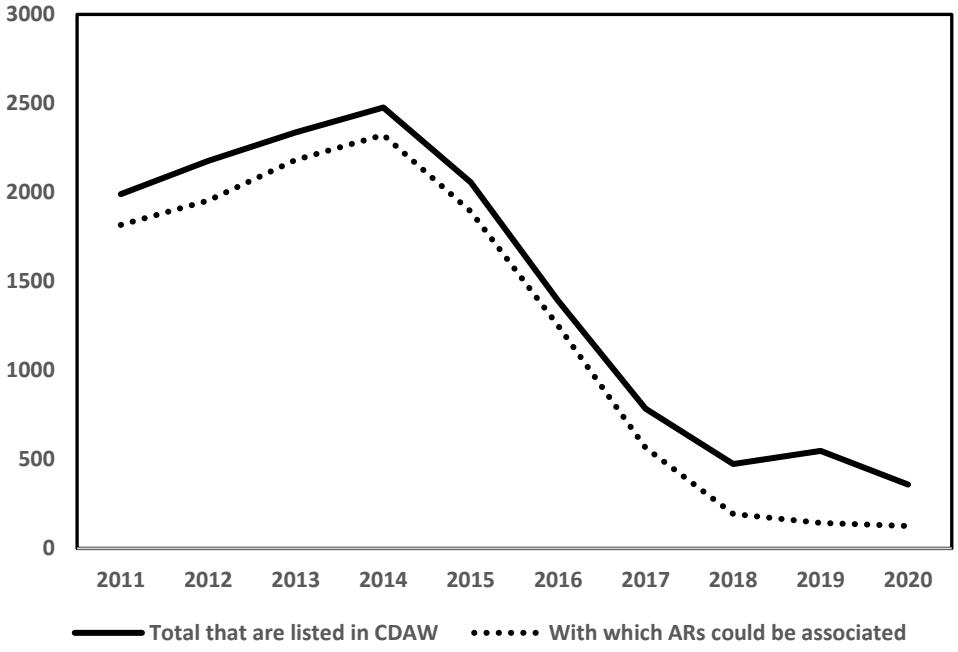

Figure 5. The variation of the total CMEs listed in CDAW and the ones for which an AR-Pool could be formed based on the criteria given in Section 2.3 and shown in Table 8 over the years from 2011 to 2020.

When we look at the data from a statistical point of view, 4895 of 14604 CMEs have been associated with their respective active region patches that makes 33.5% of the total CMEs. The results provide a huge catalogue of source HARPs with respect to CMEs if we consider that the data used for the active regions in this study comes from coronagraphs and hence the field of view (FOV) is limited as it is from the Earth's FOV. It can be said that if a HARP data from behind the seen solar surface were to be added as well to the training data simultaneously the number of these cases with associated HARPs would be around 67%.

Finally, yet importantly, the correlation between the total number of CMEs in the CDAW database and the CME cases with AR-pools associated as a result of our filtration process with respect to years is 99.8%. This can be observed in Fig. 5 in a graphical form. And the correlation between the total number of predictions made by the LSTM and the cases with one associated HARP with its respective CMEs is 97.9%. Fig. 6 depicts this correlation. This shows that the variation in the number of cases selected through the filtration process over the years (2011-2020) is similar to the actual cases in CDAW database and the variation in the model prediction for the same period is also similar. This result indicates that the produced model has in fact generalized the problem and was able to learn the patterns for the onset of the CMEs. The correlation of the predictions where a source AR was selected from the AR-

pool with the sunspot number from the Sunspot Index and Long-term Solar Observations (SILSO) and the American Association of Variable Star Observers (AAVSO) organizations is 98.1% and 98.5%, respectively. This shows that the produced results are consistent with the solar cycle as well. Although the solar cycle is not graphed in Fig. 6 the recent solar maximum can be observed in the figure at the start of 2014 (refer to the horizontal axis of Fig. 6), the same can also be observed in Fig. 5. This is due to the positive correlation with the solar cycle. A detail of these correlations with individual parameters is listed in Table 10. Aggregated annual values of the parameters of associated events were calculated before checking these correlations.

Parameter	Correlation With SILSO	Correlation With AAVSO	Mean Correlation
Linearspeed	0.96	0.97	0.97
Second order initial speed	0.95	0.96	0.96
Second order final speed	0.96	0.97	0.97
Second order speed at 20R	0.96	0.97	0.96
Mass	0.86	0.87	0.87
KE	0.79	0.81	0.80
Momentum	0.83	0.85	0.84
MEANJZD	0.85	0.87	0.86
TOTUSJZ	0.94	0.95	0.94
SAVNCPP	0.81	0.82	0.81

Table 10. The correlations between the parameters of the CME events associated as a result of this study with each other and the sunspot number from SILSO and AAVSO organizations.

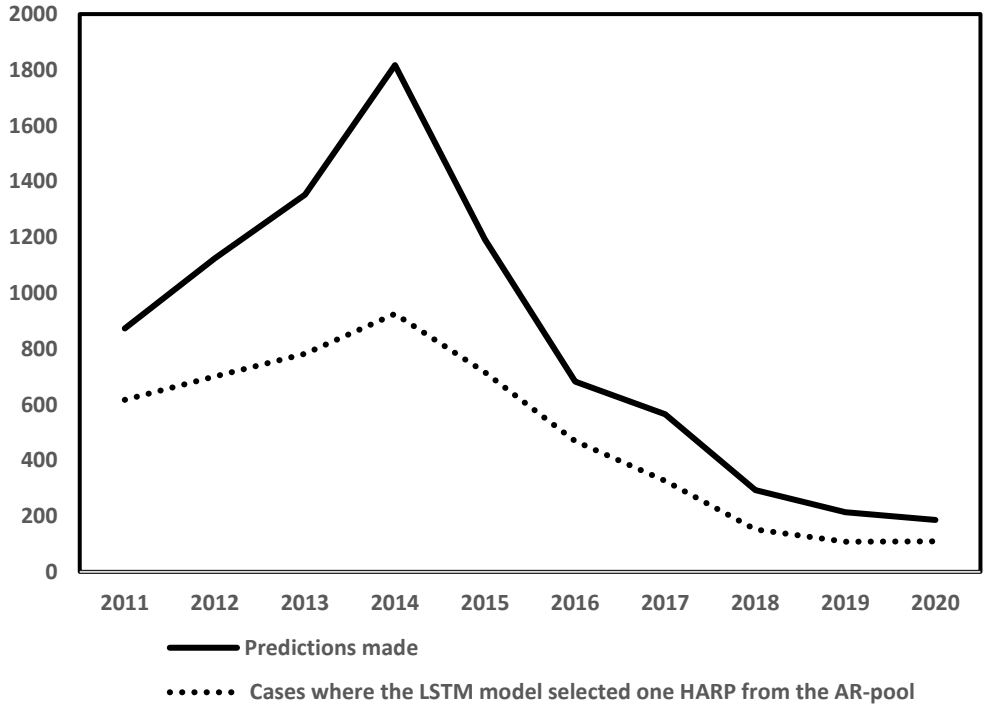

Figure 6. The variation of the total predictions made by the LSTM model and the cases where the model selected one HARP from the AR-pool of the CME also given in Table 9.

Table 11 shows the correlations between the CME parameters and the parameters of their associated source regions aggregated annually. These correlations are all above 0.80 and the highest is 0.96. This shows that the LSTM model used for the predictions is consistent in its predictions and does not produce random results.

CME Parameters	**Source Region Parameters**		
	MEANJZD	TOTUSJZ	SAVNCPP
Linearspeed	0.95	0.96	0.87
Second order initial speed	0.95	0.96	0.87
second order final speed	0.94	0.96	0.87
Second order speed at 20R distance	0.92	0.95	0.87
Mass	0.96	0.88	0.81
KE	0.96	0.87	0.80
Momentum	0.96	0.87	0.80

Table 11. The correlations between the parameters of CMEs and those of their respective source regions as a function of their annual total number.

The DONKI database only lists 120 source ARs for the CMEs. We believe the technique shown in this study can accelerate the studies regarding the investigation of onset of CMEs. This database provides a unique opportunity to study the triggering mechanism for the onset of the CME and its identification as any study related to this kind of investigation fundamentally needs a large dataset. A later study is planned to further investigate the relationships between different magnetic parameters of HARPs to study the mechanism for the onset of CMEs based on this database.

Acknowledgement:

This paper was supported by The Scientific and Technological Research Council of Turkey (TUBITAK) project numbered 117F336. This study made use of the CDAW database which is generated and maintained at the CDAW Data Center by NASA and The Catholic University of America in cooperation with the Naval Research Laboratory. Data used in this study came from SOHO which is a project of international cooperation between ESA and NASA. This research used version 2.1.2 of the SunPy open-source software package.

Data Statement

Data are available in a repository and can be accessed via https://doi.org/10.6084/m9.figshare.14512860.v3

References:

Bobra, M.G. and Ilonidis, S., 2016, ApJ (the), 821, 127.

Cavus, H., Araz, G., Coban, G.C., Raheem, A. and Karafistan, A.I., 2020, Adv. in Space Res. 65, 1035-1047.

Chandra, R., Gupta, G. R., Mulay, S., & Tripathi, D., 2015, MNRAS, 446, 3741–3748.

Falconer, D.A., Moore, R.L. and Gary, G.A., 2003, J. of Geophysical Res.: Space Physics, 108.

Gopalswamy, N., Yashiro, S., Michalek, G., Stenborg, G., Vourlidas, A., Freeland, S., Howard, R., 2009, Earth, Moon, and Planets, 104, 295-313.

Gosling, J.T., 1993, J. of Geophysical Res.: Space Phys., 98(A11), 18937-18949.

Green, L., 2016, 15 Million Degrees: A Journey to the Centre of the Sun. Penguin UK, 212

Hudson, H. and Ryan, J., 1995, Annual Review of A&A, 33, 239-282.

Inceoglu, F., Jeppesen, J.H., Kongstad, P., Marcano, N.J.H., Jacobsen, R.H. and Karoff, C., 2018, ApJ (the), 861, 128.

Karim, F., Majumdar, S. and Darabi, H., 2019, IEEE Access, 7, 67718-67725

Kingma, D.P. and Ba, J., 2014, preprint (arXiv:1412.6980)

Leka, K.D. and Barnes, G., 2003, ApJ (the), 595, 1277. (a)

Leka, K.D. and Barnes, G., 2003, ApJ (the), 595, 1296. (b)

Liu, H., Liu, C., Wang, J.T.L. and Wang, H., 2019, ApJ (the), 877, 121.

Liu, H., Liu, C., Wang, J.T.L. and Wang, H., 2020, ApJ (the), 890, 12.

Manoharan, P. K., 2006, Sol. Phys. 235, 345–368.

SunPy Community, Mumford, S. J., et al. 2015, Comput. Sci. Discov., 8, 014009

[dataset] Raheem A, Cavus H, Coban G. C., Kinaci A. C., Wang H., T. L. Wang, J. 2021, An investigation of the causal relationship between sunspot groups and coronal mass ejections by determining source active regions. figshare. https://doi.org/10.6084/m9.figshare.14512860.v3

Richardson, I.G., 2014, Sol. Phys., 289, 3843-3894.

Schrijver, C.J., 2007, ApJ Letters, 655, L117.

Toriumi, S. and Wang, H., 2019, Living Reviews in Solar Physics, 16, 1-128.

von Forstner, J.L.F., et al, 2018, J. of Geophysical Res.: Space Phys., 123, 39-56.

Wang, J., Zhang, J., Deng, Y., Li, J., Tian, L. and Yang, X., 2002, Sci. China Ser. A-Math. 45, 57–64.

Yan, X.-L., Qu, Z.-Q., & Kong, D.-F., 2011, MNRAS, 414, 2803–2811.

Links

URL-1 https://cdaw.gsfc.nasa.gov/CME_list/index.html

URL-2 http://jsoc.stanford.edu/doc/data/hmi/sharp/sharp.htm

URL-3 https://kauai.ccmc.gsfc.nasa.gov/DONKI/

URL-4 https://doi.org/10.6084/m9.figshare.14512860.v3